

https://dspace.org/

Recombinant expression and bioinformatic analysis of the Hepatitis B virus X protein

Thompson, Liam Jed

2012-09-18
http://hdl.handle.net/10539/11962

Downloaded from a DSpace repository. https://dspace.org/

RECOMBINANT EXPRESSION AND BIOINFORMATIC ANALYSIS OF THE HEPATITIS B VIRUS X PROTEIN

LIAM JED THOMPSON

A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy

Johannesburg, May 2012

Declaration

I declare that this thesis is my own, unaided work. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

Signature:

May 15, 2012

Abstract

There are an estimated 350 million people chronically infected with Hepatitis B Virus (HBV), of which approximately 600 000 die each year from HBV complications including cirrhosis and liver cancer. The X protein from HBV (HBx) has been implicated in the progression of chronic HBV to liver cancer and has been reported to manipulate several critical cellular pathways. These include the cell cycle, the tumour suppressor protein p53, protein degradation and signal transduction pathways. The role of these interactions in HBV replication and the viral lifecycle is currently unknown. The lack of animal models and infectable cell lines together with solubility and stability issues related to the HBx protein have made progress difficult. The reliance on approximate cellular and animal models has yielded many discordant studies that have confounded our interpretations of the role of HBx. There have been no novel approaches attempting to express HBx at a quantity and quality sufficient for high resolution X-ray and nuclear magnetic resonance structural determination. Additionally no bioinformatic analyses have been applied to HBx, and thus distinctive features of HBx that may be responsible for these challenges have not been reported.

This thesis describes the detailed experimentation to express and purify HBx in a functional, soluble and stable form. The study focussed on *Saccharomyces cerevisiae* and Semliki Forest Virus (SFV) expression systems, together with the use of a solubility-enhancing Maltose Binding Protein protein tag (MBP). The *S. cerevisiae*-based pYES2 and YEp and mammalian expression vectors showed production of HBx protein. However HBx that had been expressed using *S. cerevisiae* and human cells could not be reliably detected in Western blots using antibodies raised against *E. coli*-expressed HBx. This result was despite the positive visualisation of HBx using the same antibodies and immunofluorescence microscopy. This validated previous reports describing the variable antigenicity of HBx. Furthermore these findings supported the decision to develop eukaryotic-based HBx expression vectors as results suggested structural differences between eukaryote and prokaryote expressed protein. HBx was subsequently detected and purified in a soluble and active form using an MBP tag as well as a SFV expression vector. All of these options provide an excellent point from which further work at optimising HBx expression and structural elucidation can occur.

Bioinformatic analysis of HBx suggested the presence of protein disorder and protease sensitive sites within the negative regulatory domain of HBx. Literature descriptions of the molecular promis-

cuity that protein disorder allows, offers an explanation for the presence of the discordant findings on HBx interactions and functions. It is generally accepted that proteins containing disorder are tightly regulated and thus experimental systems employing overexpression methodologies may encourage cellular toxicity and non-specific interactions through the use of short linear motifs. Evolutionary analysis of HBx sequences revealed that the eight HBV genotypes (A-H) showed concordance regarding synonymous and non-synonymous substitutions at the overlapping and non-overlapping domains of *hbx*. Substitutions in *hbx* were most common at positions where a synonymous substitution occurred in the overlapping partner gene. The presence of sites under positive, neutral and negative selection were identified across the length of HBx. The different genotypes showed positive selection indicating selective pressures unique to each, thus offering a contributing explanation for the variable disease severity observed between the subtypes.

Overall, this thesis has provided novel methods to express and purify HBx in *S. cerevisiae* and mammalian cells. These methods, together with an increased understanding of the nature of HBx sequences through bioinformatic analysis, pave the way to conduct both structural studies and biological assays to elucidate the genuine roles of HBx in the HBV lifecycle and its contribution to the progression to liver cancer.

Acknowledgements

Professor Patrick Arbuthnot, for giving me the freedom to pursue my ideas in the lab and allowing me to find a way out of my naivety by myself, it will serve me well in years to come.

Dr Wolfgang Prinz, firstly for generating the HBx antibodies, and secondly for keeping the laboratory interesting and amusing even when HBx was getting the better of my mood. Say no more, say no more.

Dr Pierre Durand, for your friendship and your help with the bioinformatics, you were a life saver.

To my wife, Samantha, for supporting me through a lengthy project that yielded little in the way of joy. Listening to my aggravations and offering solid advice carried me through to the end. I am very fortunate to be able to share life with someone like you.

My parents and family, for pretending to know what I am working on and always being concerned about my progress. You can finally stop asking.

Lastly, I would like to thank the following funding bodies for their financial contributions during my studies. The University of the Witwatersrand, CANSA for funding the project, the National Research Foundation, the Poliomyelitis Research Foundation, Stella and Paul Loewenstein and the German Academic Exchange (DAAD). Patrick, for personal funding when everything else dried up.

Preface

Aspects of work described in this thesis have been published or presented at conferences.

Journal Articles

1. P. Arbuthnot, **L. J. Thompson** (2008) Harnessing the RNA interference pathway to advance treatment and prevention of hepatocellular carcinoma. World Journal of Gastroenterology. Vol 14, 11. p1670-81. http://dx.doi.org/10.3748/wjg.14.1670

2. **L. J. Thompson**, P. Arbuthnot. Evolutionary analysis of the Hepatitis B Virus X Protein. Manuscript in Preparation.

Conference Proceedings

1. **L. J. Thompson**, P. Arbuthnot (2010) Signalling aptamers for the detection of HBV serological markers in patients at risk of hepatocellular carcinoma. EMBO Course - Microscopy: From genome scale to the single molecule. Pretoria, South Africa.

2. **L. J. Thompson**, P. Arbuthnot (2010) Signalling aptamers for the detection of HBV serological markers in patients at risk of hepatocellular carcinoma. South African Society for Biochemistry and Molecular Biology. 18 - 20 January 2010. Bloemfontein, South Africa.

3. **L. J. Thompson**, P. Arbuthot (2009) Signalling aptamers for the detection of HBV serological markers in patients at risk of hepatocellular carcinoma. 21st IUBMB- 12th FAOBMB International Congress of Biochemistry and Molecular Biology & Young Scientists Program. 30 July - 2 August 2009. Shanghai, China.

4. **L. J. Thompson**, P. Arbuthnot (2008) Signalling aptamers for the detection of HBV serological markers in patients at risk of hepatocellular carcinoma. IUBMB/EMBO/FEBS Advanced Summerschool Africa - Molecular and Cellular Basis of Infection. Hermanus, South Africa.

Contents

Abstract iii

Acknowledgements v

Preface vi

List of Figures xii

List of Tables xiv

1 Introduction 1

1.1 HBV Epidemiology . . . 2

1.1.1 Infection with HBV and Pathogenesis . . . 2

1.2 HBV and Hepatocellular Carcinoma . . . 3

1.2.1 HBV Genome Organisation . . . 6

1.2.1.1 Core Protein . . . 7

1.2.1.2 Surface Proteins . . . 8

1.2.1.3 Polymerase Protein . . . 8

1.2.2 HBV X Protein . . . 8

1.3 Systems of Protein Expression . . . 12

1.4 Bioinformatic Analysis of HBx . . . 13

1.5 Aims of Research . . . 14

2 Exposition on the Expression, Purification and Detection of Recombinant HBx 16

2.1 Introduction . . . 16

2.2 Materials and Methods . . . 18

2.2.1 Bacterial Expression Vector Construction . . . 18

2.2.1.1 pET Vectors . . . 18

2.2.2 Codon Optimisation . . . 19

2.2.3 Yeast Expression Vector Construction . . . 19

2.2.3.1 pYES2 Expression Vectors . . . 19

2.2.3.2 YEp Expression Vectors . . . 19
2.2.4 Mammalian HBx Expression Vector Construction . . . 21
2.2.4.1 pCI-Neo . . . 21
2.2.4.2 pCEP4 Expression Vector . . . 21
2.2.5 PCR Conditions . . . 21
2.2.6 Protein Induction, Purification and Detection . . . 22
2.2.6.1 Expression of HBx in *E. coli* pET15b and pET-ELP Vectors . . . 22
2.2.6.2 Expression of HBx in *S. cerevisiae* pYES2 and YEp Vectors . . . 22
2.2.6.3 Expression of HBx in Mammalian Vectors . . . 24
2.2.6.4 Protein Concentration from Growth Medium Supernatant . . . 24
2.2.6.5 6xHIS tagged proteins . . . 24
2.2.6.6 Western blotting . . . 25
2.2.6.7 Mass Spectrometry Protein Identification . . . 25
2.2.7 Immunofluorescence Microscopy Imaging . . . 25
2.2.7.1 Immunofluorescence Sample Preparation . . . 25
2.2.7.2 Antibody Staining of Fixed Cells . . . 26
2.2.8 HBx mRNA detection using RT-PCR . . . 26
2.2.9 Protein Disorder and Protease Susceptibility Prediction . . . 27
2.3 Results . . . 28
2.3.1 pET-ELPI-HBx protein purification . . . 28
2.3.2 Evaluation of *S. cerevisiae* Expression Vectors . . . 29
2.3.2.1 pYES2 vectors . . . 29
2.3.2.2 YEp vectors . . . 29
2.3.3 Western blot Analysis . . . 32
2.3.4 Detection of HBx mRNA in induced cultures and cell lines . . . 32
2.3.4.1 GFP-Intein fusion vectors . . . 33
2.3.5 Mass Spectrometry . . . 33
2.3.6 Immunofluorescence detection of HBx protein in induced yeast cultures and mammalian cell lines . . . 34
2.3.7 Protease Sensitivity and Disorder Prediction . . . 44
2.4 Discussion . . . 47

3 Expression and Purification of HBx Protein using a combination of Maltose Binding Protein Tag and Semliki Forest Virus 52

3.1 Introduction . . . 52
3.2 Materials and Methods . . . 54

3.2.1 pMAL-HBx Vector Construction . . . 54
3.2.2 Semliki Forest Virus Expression Vectors . . . 54
3.2.2.1 pSFV-b12a vector . . . 54
3.2.3 Protein Induction and Purification . . . 55
3.2.3.1 pMAL-c2 Vectors . . . 55
3.2.3.2 Cleavage of MBP-HBx using Factor Xa . . . 57
3.2.4 *In Vitro* Transcription (IVT) and Electroporation of SFV Vectors . . . 57
3.2.5 Harvesting of Viral Particles . . . 58
3.2.6 Viral Titering by Immunofluorescence Microscopy Imaging . . . 59
3.2.7 HBx Activity Assay . . . 59
3.3 Results . . . 60
3.3.1 Expression of MBP-HBx fusion protein in *E. coli* . . . 60
3.3.2 Cleavage of MBP-HBx fusion with Factor Xa . . . 60
3.3.3 *In vitro* activity assay of HBx protein . . . 60
3.3.4 Expression of MBP-HBx using pSFV-b12a . . . 62
3.3.5 Expression of MBP-HBx using pSFV-S2-9-pac2A-CD5-MBP-HBx . . . 63
3.4 Discussion . . . 65

4 Evolutionary Analysis of Hepatitis B Virus HBx Protein 68
4.1 Introduction . . . 68
4.2 Methods and Materials . . . 71
4.2.1 Sequence Extraction and Analysis . . . 71
4.2.2 Calculation of Codon Position Substitutions using Entropy . . . 72
4.2.3 Branched Site Models . . . 72
4.2.4 K_a/K_s estimations . . . 72
4.3 Results . . . 73
4.3.1 Entropy Calculations for *hbx* Reading Frames . . . 73
4.3.2 K_a/K_s Estimates of Positive Selection . . . 73
4.3.3 Sites Under Positive Selection . . . 77
4.3.4 Branched Sites Selection . . . 81
4.3.5 Posterior Probability Estimates of Omega . . . 85
4.4 Discussion . . . 87

5 General Discussion and Conclusions 92
5.1 Approaches used to Express and Purify HBx . . . 92
5.1.1 Problems with Antibodies . . . 94
5.2 Instability, Disorder and Moonlighting in HBx . . . 94

5.3 The Evolution of HBx . . . 98
5.3.1 Evolutionary Patterning and Drug Targeting (Posterior Probability Estimates of Omega) . . . 99
5.4 Future Work and Recommendations . . . 100

Appendices **103**

A Supplementary Data **104**
A.1 Supplementary Data . . . 104
A.1.1 Testing for overlapping antigenic site reactivity of anti-HBx antibodies . . . 104
A.1.2 Immunofluorescence Microscopy Control Images . . . 104
A.1.3 Box and Whisker Key . . . 111
A.1.4 Sites Under Selection . . . 111

B Standard Laboratory Methods **114**
B.1 Bacterial growth, storage and transformations . . . 114
B.1.1 Bacterial growth and storage . . . 114
B.1.2 Preparation of chemically competent *E.coli* . . . 114
B.1.3 Bacterial transformations . . . 114
B.1.4 Transformation and growth of *S. cerevisiae* . . . 115
B.2 Plasmid DNA preparation, electrophoresis and extraction . . . 115
B.2.1 Plasmid DNA preparation . . . 115
B.2.2 Phenol/chloroform extraction of DNA . . . 116
B.2.3 Agarose gel electrophoresis . . . 116
B.2.4 DNA purification from agarose gels . . . 117
B.2.5 RNA extraction from adherent mammalian cells seeded in 10 cm dishes . . . 117
B.2.6 *S. cerevisiae* RNA extraction . . . 117
B.2.7 DNA Sequencing . . . 118
B.3 Cloning Protocols . . . 118
B.3.1 pTZ57R/T TA Cloning Protocol . . . 118
B.3.2 Preparations for Bacterial Cultures . . . 118
B.3.2.1 ZYM-5052 Media . . . 118
B.3.2.2 Luria-Bertani Broth (LB) . . . 119
B.3.2.3 Agar Media . . . 119
B.3.2.4 Antibiotic Stock Solutions . . . 119
B.3.2.5 IPTG and X-galactosidase . . . 119
B.4 Preparations for Eukaryotic Cell Cultures . . . 119

B.4.1 Maintenance and Transfection of plasmid DNA into mammalian cells 119
B.4.1.1 Antibiotic Solutions for tissue culture 119
B.5 General Solutions 120
B.5.1 DNA/RNA Electrophoresis Buffers 120
B.5.1.1 (50x) Tris-acetate- EDTA (TAE): 120
B.5.1.2 (10x) Tris-borate-EDTA (TBE): 120
B.5.1.3 DNA loading Buffer 120
B.5.1.4 RNA Loading Buffer 120
B.5.2 SDS-PAGE and Western blotting 120
B.5.2.1 Staining solution 121
B.6 General ELISA Protocol 121

C Python Computer Code **123**

Bibliography **125**

List of Figures

1.1 Illustration of the life cycle of HBV. 4
1.2 Depiction of the progression of acute and chronic HBV infection. 5
1.3 Depiction of the progression of chronic HBV infection to HCC. 7
1.4 Depiction of the major elements of the HBV genome 9
1.5 Model of interactions of HBx with cellular targets and mitochondria 11
1.6 Geographic distribution of the 8 genotypes of HBV 14

2.1 General method of purifying proteins attached to an ELP tag 23
2.2 Representative, Western blot examples of eukaryotic expression. 30
2.3 RT-PCR for full length mRNA by DNA contructs. 32
2.4 Peptides identified by mass spectrometry of GFP-Intein-HBx-6xHIS protein 33
2.5 Immunofluorescence microscopy image compilation of the non-secreted pYES2-HBx vector 35
2.6 Immunofluorescence microscopy image compilation of the 6xHIS tagged pYES2-HHBx vector 36
2.7 Immunofluorescence microscopy image compilation of the *S. cerevisiae* HBx secreting YEp-HBx vector 37
2.8 Immunofluorescence microscopy image compilation of the *S. cerevisiae*, secreting, HIS tagged, YEp-HHBx vector 38
2.9 Immunofluorescence microscopy image compilation of the pCEP4-GFP-I-HBx fusion vector. 39
2.10 Immunofluorescence microscopy image compilation of the wildtype pCI-HBx vector. . 39
2.11 Immunofluorescence microscopy image compilation of the codon optimised pCI-6xHIS-HBX vector. 40
2.12 Immunofluorescence microscopy image compilation of the yeast fusion protein YEp-GFP-I-HBx-6xHIS vector 42
2.13 Immunofluorescence microscopy image compilation of the yeast fusion pYES2-GFP-I-HBx-6xHIS vector 43
2.14 Plots showing predicted disorder regions over the sequence length of HBx 45

2.15 Predicted PEST and D-BOX sequences of HBx 46
2.16 CD spectrograph of soluble HBx purified from pMAL expressed fusion protein. 46

3.1 Schematic showing cloning of the MBP-HBx construct into SFV vectors 56
3.2 Affinity elution profile of *E.coli* expressed MBP-HBx 60
3.3 G-75 Sephadex elution profile of *E.coli* expressed MBP-HBx, post cleavage with Factor Xa 61
3.4 Western blot and Coomassie stain of MBP-HBx and time cleavage assay of HBx from MBP 61
3.5 Comparative fluorescence intensities with standard deviations of different wells of IL8-GFP containing cells after exposure to either MBP, TNFα, MBP-HBx or HBx 62
3.6 Western blot of cell and supernatant protein samples of the pSFV-b12A-ppt-MBP-HBx vector 63
3.7 Western blot from pSFV-S2-9-pac2A-CD5-MBP-HBx 64

4.1 Illustration of the overlapping genes and relative reading frames of HBV 69
4.2 Cumulative entropy over the three reading frames of HBx. 74
4.3 Comparison between the K_a/K_s ratios of HBx from the genotypes A-F. 76
4.4 K_a/K_s values plotted over the length of HBx. 77
4.5 Sites in the genome of HBV shown to be under positive selection. 80
4.6 Phylogenetic tree used to determine branched sites under selection. 83
4.7 Posterior probabilities of HBx 86

A.1 Immunofluorescence image compilation of the negative control pCI-eGFP vector . . . 105
A.2 Immunofluorescence image compilation of the mammalian negative control pCI-Neo vector 106
A.3 Immunofluorescence image compilation of the HBx negative, yeast control YEp-eGFP vector 107
A.4 Immunofluorescence image compilation of the negative, empty yeast control YEpHF . 108
A.5 Immunofluorescence image compilation of the negative, empty yeast control pYES2 vector 109
A.6 Immunofluorescence image compilation of the negative, empty yeast control pYES2-GFP vector 110
A.7 Interpretive key for box and whisker plots 111

List of Tables

2.1 Summary of different HBx expressing clones, experimental conditions used and results obtained. 31

3.1 Components of a standard IVT reaction for the generation of capped SFV mRNA. . . 58

4.1 Summary of Log-likelihood values (lnL) and parameter estimates for models applied to HBx across all genotypes of HBV using PAML Models 0, 1, 2, 7 and 8. 79

4.2 Likelihood ratio tests (LRT) comparing the lnL values of PAML Models 1 versus 2 and Models 7 versus 8. 81

4.3 Branched site model parameter estimates for HBx sequences across genotypes A-H. 84

A.1 Sites in HBx under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. 111

A.2 Sites in HBc under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. 112

A.3 Sites in HBs under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. 112

A.4 Sites in HBp under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. 113

Chapter 1

Introduction

In 1947 the British doctor, F. O. MacCallum suggested that a virus was the aetiological agent of a disease called homologous serum hepatitis[1]. Following a series of observations on soldiers who received a Yellow Fever vaccine, MacCallum coined the term Hepatitis A and B for the forms of the disease spread primarily through contaminated food and blood respectively. It was only in 1962 that Baruch Blumberg and Harvey Alter discovered antigenic evidence of Hepatitis B whilst trying to develop screens to determine genetic differences in human populations[2]. This antigen, initially called the Australia antigen was associated with clinical hepatitis[3], and was later discovered to be the surface antigen of the Hepatitis B Virus (HBV). Electron microscopy studies subsequently revealed the presence of spherical particles (42nm in diameter particles), now known as Dane particles which are the infectious HBV virions[4].

There are two outcomes to HBV infection, either an acute infection where the disease is resolved within 6 months of infection, or a chronic infection where the HBV virus persists indefinitely after infection. There is a strong correlation between chronic infection and the risk of developing either cirrhosis and/or hepatocellular carcinoma (HCC)[5]. HCC is amongst the most prevalent tumours in the world's populous regions and has a mortality rate equal to its annual incidence. HCC has a very rapid progression and poor outcomes. The surface antigen of HBV (HBs) serves as a diagnostic marker for HBV infection but is not a marker for HCC development. Currently serum alphafetoprotein (AFP) levels serve as a predictive marker for HCC as persistently elevated AFP levels ($>$200 ng/mL) are a strong predictor of HCC presence[6]. The sensitivity of AFP is however only 60 %, meaning that elevated levels are only observed in 60 % of patients. Thus there is a need for specific markers to diagnose HBV-related HCC and it may be possible to utilise another protein from HBV as a surrogate. It is of primary importance to understand the role of HBV virulence factors in HCC development and a diagnostic based on one or more of these factors would be very helpful. One such protein is the X protein (HBx) which has been suggested in a multitude of studies to be involved in the progression of chronic HBV infection to HCC. However studies have not proved unequivocally that HBx is involved

in disease progression and there are many conflicting reports. The structural determination of HBx would assist in ours understanding of the role of HBx, however studies thus far have focussed exclusively on expressing HBx in *Escherichia coli*. As HBx is an unstable, poorly immunogenic and insoluble protein when expressed in *E. coli*, the description of evolutionary, bioinformatic and expression data in eukaryotic cells could help achieve this goal.

In this thesis the applicability of expressing HBx in *Saccharomyces cerevisiae* and mammalian cells was assessed together with the elucidation of distinguishing sequence characteristics through a bioinformatic approach. A variety of plasmid vectors were constructed and tested for the production of HBx. Furthermore several predictive bioinformatic models and algorithms were applied to the HBx sequence to aid in the interpretation of interaction and protein expression data.

1.1 HBV Epidemiology

Approximately 40 % of the world's population has either come into contact with, or are carriers of HBV. In 2005 this amounted to some 350 million chronic infections, despite 85 % of countries having instituted universal childhood vaccination policies[7]. In 2000 it was estimated that 620 000 deaths occurred per year as a result of complications related to chronic HBV infection including HCC and/or cirrhosis[8]. Furthermore, countries which have not implemented vaccination policies continue to have high prevalence rates and significant disease burdens.

Worldwide, there is a large difference in HBV prevalence rates, ranging between 0.1 % and 20 %. Typically, low prevalence areas lie in the developed economies of the West (Europe, North America and Australasia), whereas the high prevalence areas are in South-East Asia, China and sub-Saharan Africa. In the latter, most infections occur at a young age ($<$5 years) and the chance of progressing to a chronic infection is inversely proportional to the age at infection. Perinatal infection carries a chance of chronicity of almost 90 %[9], whereas adult infections only carry a 5 % chance[10]. Together with HIV, TB and malaria, HBV contributes to the excessive disease burden in sub-Saharan African countries and places significant strain on public health infrastructure.

1.1.1 Infection with HBV and Pathogenesis

Upon infection with the virus, the site for replication and persistence is in the liver. This liver tropism is a commonality between all the members of the hepadnavirus family although the exact mechanism and receptor(s) used to gain entry into hepatocytes is not known. As illustrated in Fig 1.1, upon entry the partially double stranded HBV genome is delivered to the nucleus and converted by host enzymes into an episomal covalently closed circular DNA (cccDNA) intermediate for viral replication. The cccDNA serves as a template for the transcription of the four viral open reading frames (ORFs): the 0.9 kb RNA serves as a template for HBx translation, the 2.1 kb RNA as tem-

plate for the pre-S2 (M) and S proteins whilst the 2.4 kb RNA is a template for the pre-S1 protein (L). The pregenomic RNA (pgRNA) has a greater-than-genome length and serves as a template for core (HBc) and polymerase (HBp) translation and for progeny viral genomes. HBV expression seems to be largely regulated by liver specific transcription factors as many have binding sites within the HBV genome (reviewed in[11]). The HBc and HBp proteins and the pgRNA associate with each other and form immature nucleocapsids in the cytoplasm of the infected cell. The nucleocapsids are shuttled through the endoplasmic reticulum (ER) where they associate with high concentrations of HBs proteins. The nascent viruses then bud from the ER membrane and are secreted from the cell. This is only a brief overview of the HBV lifecycle and further details can be found in the following reviews[12,13].

HBV typically incubates for an average of 90 days (range of 60 - 150 days) before symptoms start to present as illustrated in Fig 1.2-Panel A. The HBs protein is normally detected at on average at 4 weeks post infection (range of 1 - 9 weeks). However HBs may not be detected if a person is newly infected, or if the host immune system mounts a successful response to HBs. In such cases, the detection of host antibodies against HBc are used as a diagnostic marker and may provide the only evidence of disease. Following the appearance of the HBs antigen is HBe. The exact function of this protein is unknown, but it is suspected to play a role in immune system suppression.

Following immune clearance of HBs, and the emergence of antibodies against HBe, there is typically a dramatic decrease in viral replication. In a classic acute infection, the infected person will clear the HBs to undetectable levels, and antibodies will arise against HBc and HBs proteins. However, in some individuals, the HBs is not successfully cleared, and after six months the individual becomes a chronic HBV carrier[14]. A classic example of progression to chronic infection is shown in Fig 1.2-Panel B.

Although it is known that cytotoxic T cell and natural killer cell responses are needed to clear an HBV infection completely[16–18], the reasons why an acute infection becomes chronic are still not known. It is most likely that those individuals who are immuno-compromised or have a poorly developed immune system (e.g. young children) fail to mount a strong innate response and thus do not eliminate the infection[19]. Such individuals are thereafter at an elevated risk of progressing to HCC later on in life.

1.2 HBV and Hepatocellular Carcinoma

Epidemiological evidence has established a strong connection between chronic HBV infection and HCC[20,21], with a lifetime risk of developing liver cancer ranging from 5 % to 20 %. After the discovery of HBV genomic material in DNA isolated from HCC samples[22] and cell lines[23] it was suggested that HBV acts as a proto-oncogene activator due to insertional mutagenesis.

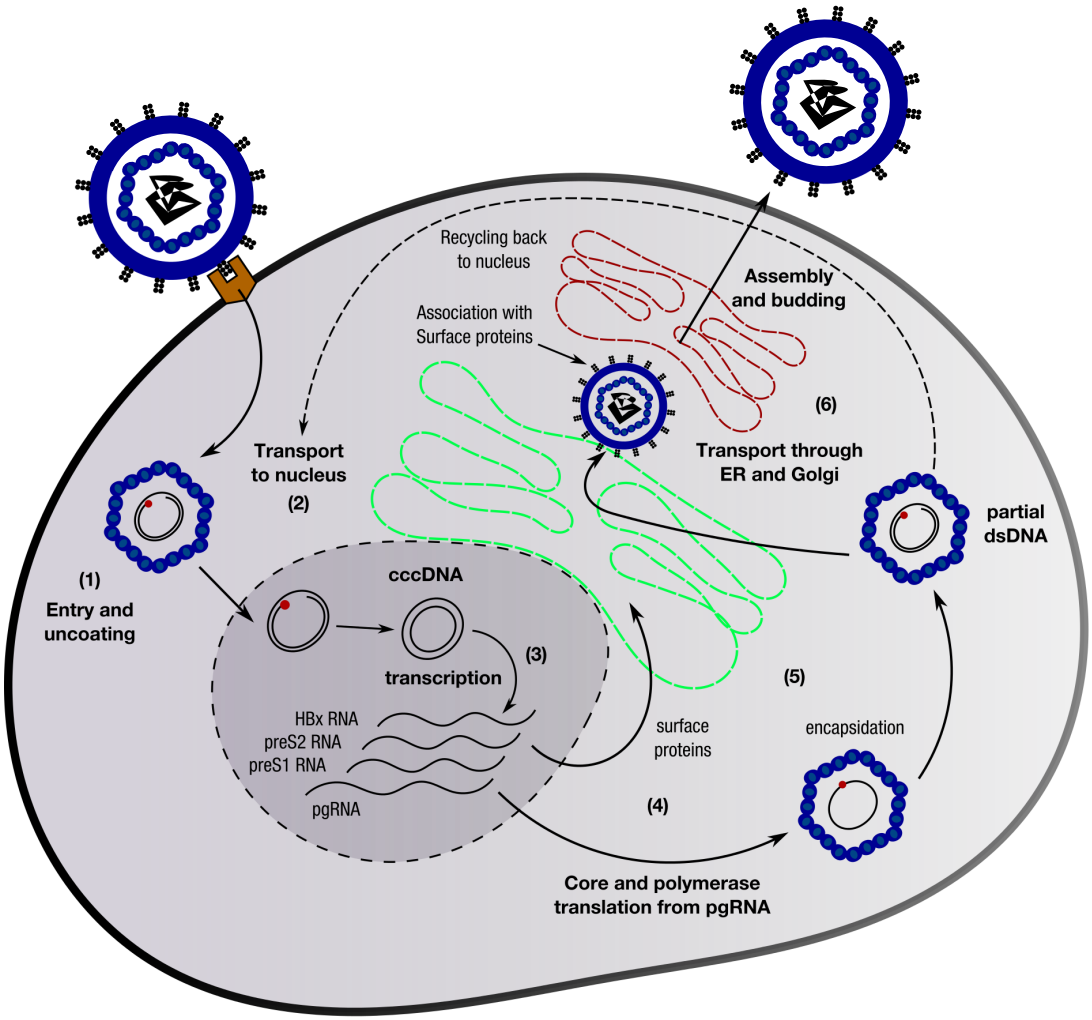

Figure 1.1: Illustration of the lifecycle of HBV from infection of a hepatocyte to the secretion of progeny viruses. 1) The virus docks with an unknown receptor and enters the cell after which it is uncoated. 2) The viral genome is transported into the nucleus and converted to cccDNA by host enzymes. 3) Transcription of the four ORFs occurs from the cccDNA template. These include hbx, preS2, and preS1/S mRNAs as well as pgRNA. The viral transcripts are exported from the nucleus, and 4) translation of the respective proteins occurs in the cytoplasm. 5) Together with the pgRNA, the polymerase (HBp) and core (HBc) proteins form immature, RNA packaged nucleocapsids. The pgRNA is converted by HBp to DNA via reverse transcription. 6) The nucleocapsids then associate with surface proteins (HBs) in the ER and Golgi bodies and are enveloped after budding from the ER. The virions are secreted from the cell.

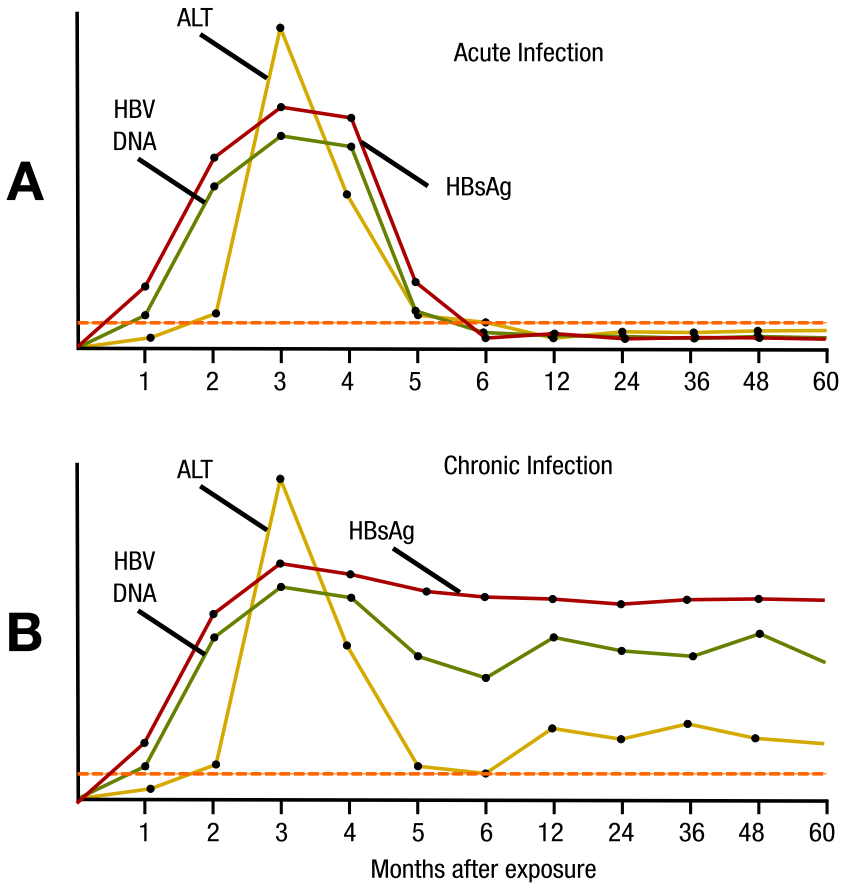

Figure 1.2: Representative timecourses of acute (Panel A) and chronic (Panel B) HBV infections showing viral antigen and DNA levels[15]. Both acute and chronic infections proceed in a similar manner, with a rapid rise in the levels of detected HBV DNA (green line), HBs (red line) and serum Alanine transaminase levels (ALT) (yellow line) during the first three months of infection. The ALT is a measurement of liver function. Thereafter acute infections are generally resolved within a further 3 months as measured by a drop in HBV DNA, HBs and ALT levels. In contrast chronic infections continue to show elevated levels of HBs and HBV DNA although ALT levels may be reduced. The dotted orange line represents an arbitrary limit of detection in both panels.

The woodchuck hepatitis virus (WHV) was the first mammalian and avian hepatitis virus described after HBV. Woodchucks chronically infected with WHV proceed to severe hepatitis and HCC in the first 2 - 4 years of life. Furthermore the presentation of woodchuck HCC is very similar to HBV-related HCC, making this a useful animal model for the study of HBV. Assessing the pattern of insertional mutagenesis in woodchuck HCC revealed that nearly all WHV genomic insertions are present near either the N-*myc*2[24], c-*myc*[24] or N-*myc*1 loci[25], resulting in their activation. Unfortunately, the same has not been observed for HBV and random integration sites were observed for HBV tumours, with only a few cases of integration occurring near an oncogene[26].

As the search for predictable integration sites of HBV remains unsuccessful, there are two other proposed models to help explain HBV-related HCC. The first model involves tumour suppressor genes in the development of HCC and is discussed below, whereas the second model proposes the presence of an oncogene within HBV and is discussed in Section 1.2.2.

The first model is based on HCC being a multistep disease process similar to that described for colon cancer (see Fig 1.3). This was reasoned because a variety of lesions were visible in people with HCC, including dysplastic nodules, altered hepatic foci and sites of HCC displaying different degrees of cellular differentiation[5]. Unlike colon cancer, the molecular characterisation of liver cancer is still being determined but a major focus has been on the tumour suppressor protein p53. This focus developed when two reports detailed that approximately half the HBV related HCC samples obtained from patients in Qidong China[27] and Southern Africa[28] contained mutations within this gene. In most of the cases described, a G to T transversion was present at codon 249. Such substitutions are characteristic of exposure to chemical carcinogens such as aflatoxin B1 which are common in these geographical regions[28,29]. Liver cirrhosis is an additional risk factor for the development of HCC and occurs when hepatocytes are replaced by scar tissue. This may be caused by chronic HBV infection, Hepatitis C infection as well as alcohol abuse and fatty liver disease. In cirrhosis, liver function is impaired, is generally irreversible and individuals with this disease are at risk of dying from liver failure or progressing to HCC.

Subsequent findings have revealed that the frequency of p53 mutations in HCCs (12 - 30 %[31]) are lower than other human cancers (approximately 50 %[32]). It seems therefore that p53 mutations are not an early event that leads to HCC, but rather a late event, similar to colon and other human cancers[33]. The second model of HBV-related HCC involves an oncogene encoded within the virus. To further understand this, a brief overview of the HBV genome organisation is described below.

1.2.1 HBV Genome Organisation

As depicted in Fig 1.4, the DNA genome of HBV (and other hepadnaviruses), is partially double-stranded and ranges in size from 3.1 - 3.3 kb, depending on the genotype or isolate[34]. All of the genes overlap to some degree with at least one other HBV gene. The four open reading frames are

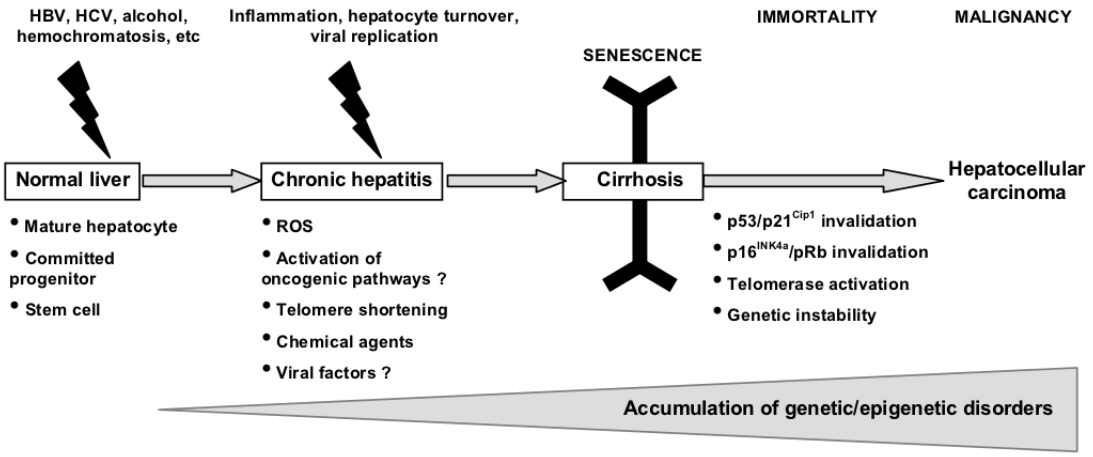

Figure 1.3: A model of the progression of chronic HBV infection to HCC as proposed by Merle and Trepo[30]. In the model, several types of liver cells could support the accumulation of genetic alterations, leading to senescence and ultimately to HCC. When the cell cycle checkpoints are overridden the cell is able to re-enter the cell cycle and telomerase activation causes DNA hypermethylation and genetic instability. The cell accumulates the changes necessary to achieve tumourigenicity and cancer stem cell capabilities.

all encoded by the DNA minus strand[35], and the three sub-genomic transcripts (2.4, 2.1 and 0.9 kb) as well as the single greater-than-genome length pgRNA, are produced during infection. The large HBs protein is translated from the 2.4 kb transcript whereas the middle and major HBs proteins are both translated from the 2.1 kb transcript. The HBx protein is translated from the 0.9 kb transcript. The pgRNA serves as the template for HBc and HBp protein translation as well as template for HBV progeny genomes. The pgRNA also contains a secondary structure ($\mathcal{E}$) signal that is present on both the 5′ and 3′ ends. The 5′ end is the one recognised by the HBV HBp and acts as the initial packaging signal[36].

1.2.1.1 Core Protein

The *core* ORF is divided into two in-frame sections, a pre-core and core domain (see Fig 1.4). Translation from the upstream pre-core codon results in the E antigen (HBe). This antigen contains a signal peptide, is proteolytically cleaved on the amino and carboxy termini, and secreted from the cell[37]. The detection of this protein is indicative of active viral replication. Translation from the downstrean core start codon results in HBc proteins which then dimerise and assemble to form viral icosahedral capsids.

1.2.1.2 Surface Proteins

During viral replication, the outer envelopes of the infectious Dane particles are added to the virus by budding from the membrane of the endoplasmic reticulum[38]. There are three viral proteins in the envelope namely S, M and L proteins. S is the smallest protein (residues 175 - 400) and defines the S domain common to all the HBs proteins. The M domain has an added pre-S2 domain (residues 120 - 174), whilst the L domain has both the pre-S1 (residue 1 - 108 or 1 - 119) and pre-S2 domains[35]. The exact size of these proteins depends on the HBV genotype. The HBs proteins are co-translated in the ER due to the presence of several transmembrane sequences. Regions in the ER which are high in the S but not M or L proteins may aggregate in the Golgi complex, bud into the lumen of the ER, and then be secreted from the cell as 22 nm particles. The pre-S1 domain, displayed on the outside of the S antigen, can elicit virus-neutralising antibodies and is thought to be partly involved in the viral receptor responsible for infection[39]. The function of the M domain is not understood and the protein is not essential for the replication of HBV in hepatocyte cultures[40].

1.2.1.3 Polymerase Protein

The ORF containing *polymerase* encodes the HBp protein, which possesses both RNA- and DNA-dependent polymerase activity required during viral replication. HBp is divided into three main domains, the primase domain on the N-terminus, the polymerase domain in the middle and the RNase H-like domain on the C-terminus[41]. The primase domain primes minus strand synthesis and is eventually covalently linked to the 5′ end of the DNA minus strand[42]. It is also important for the packaging of pgRNA as well as acting as a primer for reverse transcription[43]. The middle domain gives the protein its name and is responsible for the RNA- and DNA-dependent polymerase activity of the protein. The RNase H domain is responsible for degrading the pgRNA from the RNA-DNA intermediate created during reverse transcription[44]. As the HBc, HBs and HBp proteins were of less importance in this current body of work, only a brief description of these proteins has been provided. A more detailed description of these proteins may be found in[15].

1.2.2 HBV X Protein

The second oncogenic model described above predicts that expression of one of more of the proteins of HBV may induce a malignant transformation in tissue culture cells or animal models of HBV. Suspicion of oncogenic potential initially fell on HBx in 1988 when its mRNA transcripts were detected in the livers of infected Woodchucks[45] and human patients[46]. The detection of HBx protein in HCC liver biopsy samples from human patients has been variable, ranging between 0 %[47] and 86 %[48]. This low level of concordance between studies could be as a result of antibody detection problems because HBx is known to have a low immunogenicity. However there could be an interplay

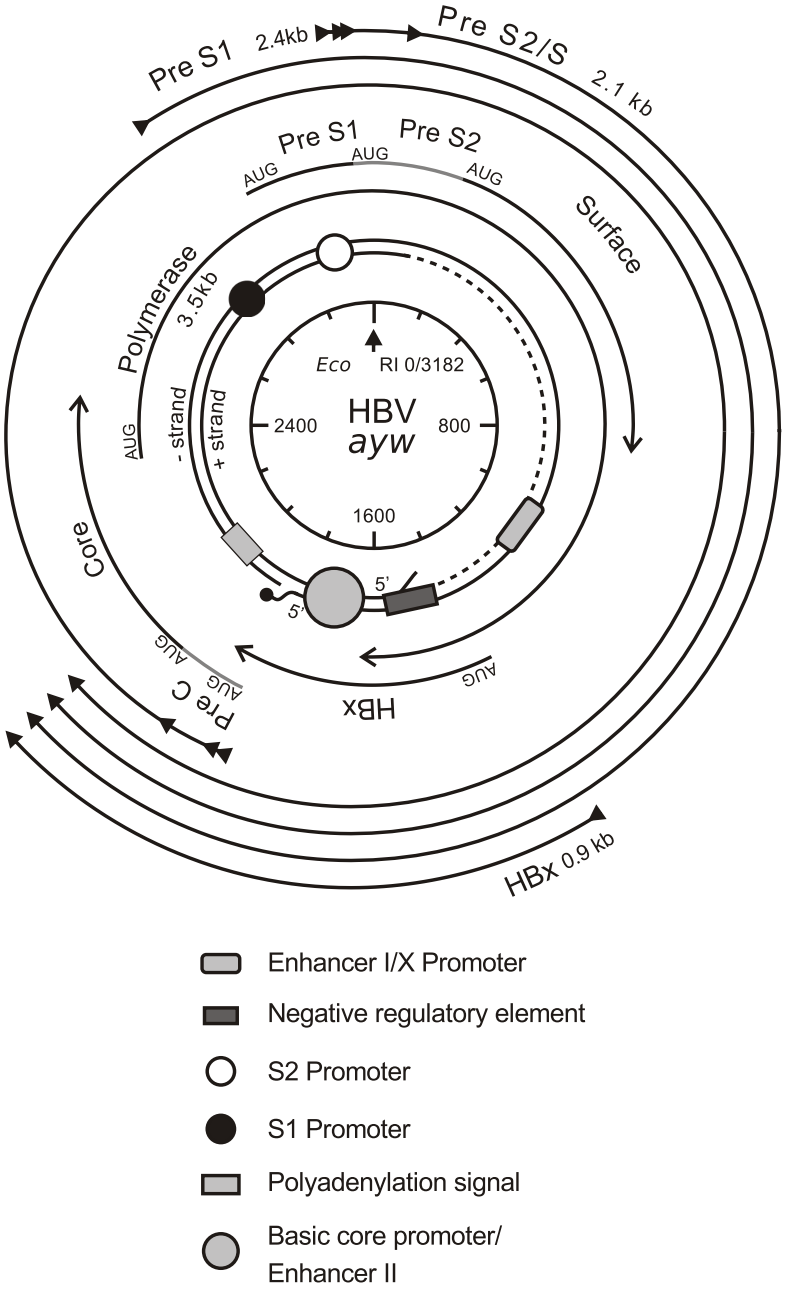

Figure 1.4: The HBV genome is a partially double stranded DNA genome with 4 open readings frames. It encodes four main genes, *polymerase*, *surface*, *core* and *hbx*. The *preS1* and *preS2* and *surface* segments are in frame with each other, and depending at which AUG codon translation occurs from, the large (S1/S2/S), middle(S2/S) and major(S) HBs proteins are produced. The *pre-core* segment is in frame with *core* and translation from this AUG results in the e antigen pre-core/core precursor. The *core* gene encodes the HBc protein which makes up the viral nucleocapsid. The *hbx* gene encodes the HBx protein which displays transactivation function.

between virus genotype, patient ethnicity, age of infection and HBx expression which has not been fully explored.

Numerous studies have used mouse models to examine the hepatocyte transformative capabilities of the HBx protein. The overexpression of HBx in CD-1 mice, which are susceptible to spontaneous hepatoma generation, showed a 10-fold increase in liver tumour frequency compared to control mice of the same strain. Changes to the liver included preneoplastic liver foci, which consisted of hepatocytes with cytoplasmic vacuolations[49]. However using mice expressing HBx at a physiologically relevant level, there was no change in the prevalence of liver tumour development[50]. Still more mouse models, observed over a period of two years, showed that at a level of HBx expression lower than that used in the CD-1 mouse study, HBx protein alone did not cause serious liver damage nor HCC[51]. These studies highlight that prudence should be shown when considering the results generated with animal models as, like the WHV-HCC progression studies discussed in section 1.2, the differences between model systems and HBV-human biology is often contrasting.

The manipulation of cell cycle and cell proliferation pathways is a major mechanism by which several tumorigenic viruses transform infected cells (e.g. SV40, HPV-16, EBV). These viruses encode non-structural proteins that can stimulate quiescent cells to enter the S phase of the cell cycle[52]. Similarly such cell cycle manipulation has been described for HBV[53]. It has been observed that serum starved cells in an arrested growth phase proceeded to the G_1 phase but not S phase when exposed to HBx[54]. Bouchard *et al* conducted a study to determine what effect HBx had on the cell cycle, both by itself and in the context of other HBV genes. Using primary rat hepatocytes, they determined that upon infection with HBV, normally quiescent hepatocytes were induced by HBx to move from the G_0 to the G_1 phase of the cell cycle[55]. The authors suggested that due to the free pool of dNTPs in quiescent hepatocytes being low, HBV (through the actions of HBx) causes the cell to move into the G_1 phase thus increasing access to dNTPs needed for viral production[56]. It has been suggested that the cell cycle is manipulated as a result of membrane depolarisation of mitochondria[57], as shown in Fig 1.5. Depolarisation results in the release of free Ca^{2+} which not only interacts with the Pyk2/FAK and SRC kinase families but is important in inducing quiescent cells to re-enter the cell cycle[58]. Of note is the stalling of the cell in the G_1 phase which may be advantageous for the virus as it then does not compete with host replication machinery, and likely prevents the initiation of apoptotic pathways brought on by premature cellular DNA synthesis and replication[55]. Furthermore as this study was conducted using physiologically relevant cells and viral levels, the data are highly suggestive of a natural function of HBx.

Most experimental evidence generated thus far suggests that HBx is not by itself directly responsible for the progression of chronic HBV to HCC. Many of the described interactions may have been a result of an inadvertent amplification of HBx protein's inherent ability to be involved in multiple

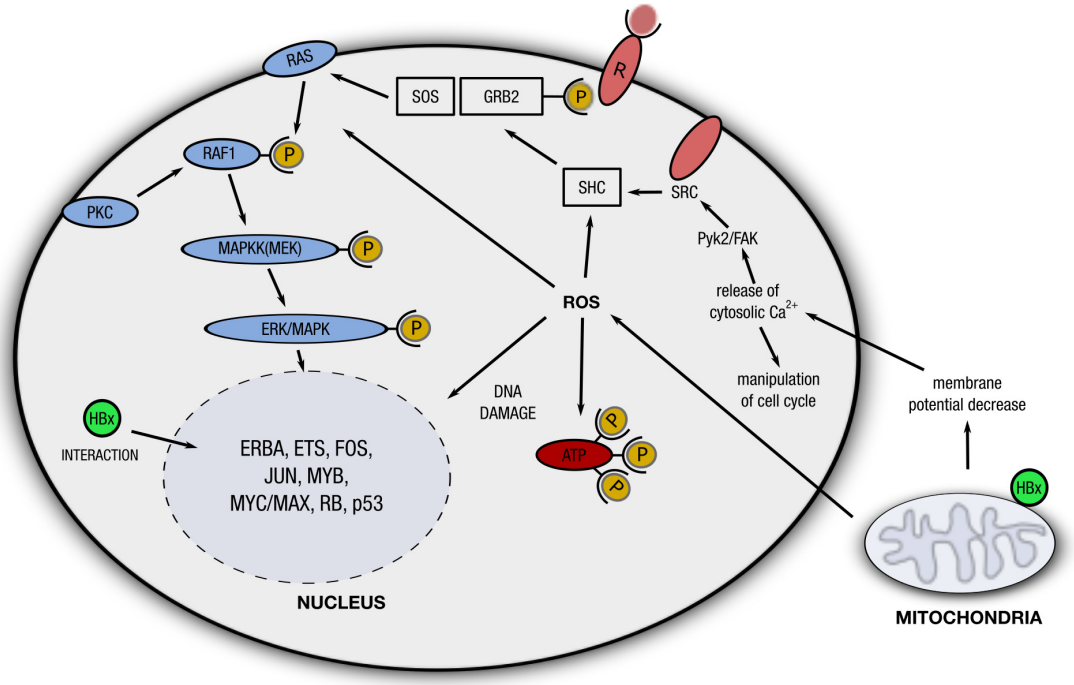

Figure 1.5: A model of the interactions of HBx with cellular targets and mitochondria (adapted from Bouchard and Schneider[59] and Katsuro and Koike[60]). Several signal transduction pathways have been shown to interact with HBx including RAS/MAPK, JNK, NF-κB, and Src (blue ovals). The most prominent interaction is with the mitochondria permeability transition pore, which results in depolarisation of the membrane potential and an increase in the cytosolic levels of Ca^{2+}[57]. This Ca^{2+} increase is involved in many cellular pathways, the most notable of which is cell cycle control. Active versions of the relevant proteins are phosphorylated as shown by a yellow circle. ROS (reactive oxygen species).

functions and interactions. It is notable that despite the increased risk of liver cancer in a chronic infection, there is little evidence to suggest that HBV is cytotoxic to liver cells under physiological conditions in an experimental setting[61–63].

It has also been common practise for HBx research to involve animal models and cell lines that do not adequately reflect human liver physiology. Studies attempting to replicate findings from other research groups have often generated different conclusions, probably as a result of different experimental conditions rather than true discrepancies. It is also possible that the use of HBx overexpression and poor non-hepatic cellular models have unintentionally ”forced” interactions between HBx and various cellular targets. It has therefore become very difficult to distil genuine versus artefactual interactions. To remedy this, greater emphasis needs to be placed on model systems that are physiologically relevant including the use of HBx in context with other HBV proteins, and at concentration of proteins observed in an infection setting.

Taken together these results and discrepancies suggest that the progression to HCC could be an interplay between HBV's manipulation of the cell cycle to enable virus production, and the immune system attempting to counter virus production through immune-mediated cell death. The ability of HBx to manipulate the cell cycle may thus set the stage for the proliferation of cells containing DNA damage leading to carcinoma formation over time. Solid data relating to the functions of HBx remain isolated and thus concrete facts about HBx functioning remain few. Continued robust explorations to uncover the true functions of HBx in HCC as well as the viral life cycle are needed, especially as the majority of studies to date have been conducted using prokaryote expressed HBV proteins.

1.3 Systems of Protein Expression

The study of proteins is often hampered by the low levels at which their genes are naturally expressed. To facilitate studies, two main protein expression systems are normally used, *Escherichia coli* and baculovirus-mediated insect cell expression. Derivations of these systems were developed many years ago, and their capabilities are increasingly insufficient for demands in producing poorly characterised proteins such as HBx.

E. coli is normally the first expression system used when generating a heterologous protein. This is facilitated by the bacteria's rapid growth cycle and dependence on inexpensive substrates. Furthermore *E. coli* genetics are well known and there are a variety of cloning vectors and mutants that one may choose from. There is however no guarantee that the chosen system will produce a correctly folded, active protein. Certainly in the case of HBx, expression with additional tags results in mostly insoluble protein located in inclusion bodies[64]. Furthermore denaturation and refolding are required to allow pure functional protein to be purified. Similarly work previously conducted in the Arbuthnot laboratory found HBx expressed from Sf9 insect cells was also insoluble and required

extensive manipulation to generate a soluble form[65]. Thus there is a significant need for additional systems of expression that can generate soluble eukaryote expressed HBV proteins.

It has long been appreciated that yeasts, and specifically *Saccharomyces cerevisiae*, are a good system for the expression of heterologous proteins (for a detailed review see[66]). Yeast are classified as a Generally Regarded as Safe (GRAS) food organism, and are acceptable hosts for the expression of pharmaceutical proteins. In contrast *E. coli* may contain cell wall endotoxins and mammalian cells are often contaminated with exogenous viruses, so products from these hosts undergo thorough testing before use. Like bacteria, yeasts can be grown on inexpensive media and they can be genetically manipulated almost as readily as *E. coli*. As yeasts are also eukaryotes, any proteins they produce are often identical to (or at least similar to) those produced in mammalian cells, whereas prokaryote produced proteins (from *E. coli*) are not. Importantly yeasts are amenable to high-level secretion of expressed proteins as well as cytosolic accumulation of soluble proteins and thus could be the ideal host to express HBV proteins and in particular HBx.

Bioinformatic analysis of HBx sequences may yield insights into the behaviour and structure of the protein and may be useful in designing an expression system. An analysis of structural disorder, the presence of protease-sensitive sites and instability may suggest reasons for the difficulty in expressing and purifying HBx.

1.4 Bioinformatic Analysis of HBx

The HBV HBp lacks 3′ - 5′ proofreading exonuclease activity resulting in an estimated misincorporation rate of 10^{-4} nucleotides[67]. Together with a replication rate capable of producing 10^{13} viral particles per day[68], the viral population accumulates an estimated 10^{9} mutations per day over the 3.2 kb HBV genome[69]. The rate is unlikely to be uniform over the length of the genome, as overlapping segments of HBV genes place evolutionary constraints on the overlapped partner gene. Thus large sequence variation accumulates in the viral population and is a source of selection for mutants that are able to resist the host immune response and to escape treatment[70].

There are eight genotypes of HBV (A-H) each differing from each other by at least 8 %[71]. Genotypes are still largely geographically separated which might allow for HBV to follow different evolutionary paths (see Fig 1.6). This is illustrated by the differences in severity, course and likelihood of developing HBV-related complications[72]. Genotypes also differ in their requirment for HBe seroconversion and the response to treatment for chronic infection[73,74]. Considering the sequence, geographic location and pathogenesis differences of the genotypes, analysis should be performed on each genotype separately to make certain that notable differences in evolutionary mechanisms are not lost due to averaging of data.

Considering the purported role of HBx in the development of HCC, and because *hbx* overlaps



Figure 1.6: Geographic distribution of the 8 declared genotypes of HBV[75]. Genotype A is largely situated in sub-Saharan and Western Africa, India and Western Europe. Genotype B is concentrated in South-East Asia and Japan. Genotype C is present in Eastern Asia, Australasia and some Pacific islands. Genotpe D is located in North Africa, Southern Europe, the Middle East, Russia and its satellite states. Genotype E is largely in central and Western Africa. Genotype F is concentrated mainly in central and South America. Genotypes G and H are also situated in central America.

with *pre-core* and *polymerase* genes, it is important to pursue studies examining the mechanism of evolution on HBx. This might offer insight into the interactions between HBx and cellular targets as well as the immune system.

Molecular evolution is part of the evolutionary process, but focussing on the DNA, RNA or protein changes that underlie organism differences cannot be inferred from morphological features. The analysis of the molecular evolution of species helps investigate the forces and mechanisms of the evolutionary process[76]. This is studied by estimating rates of nucleotide and amino acid substitutions, and then applying computational models of mutation and selection using sequence data[76]. A database of sequences serves as a collection of the evolutionary history of sequences and allows comparisons to be made that could not be done using morphological characteristics. The analysis of the molecular evolution of sequences relies on large datasets and is often very computationally intensive. To date, while HBV sequences lie within a single repository these have not been compiled nor sufficiently annotated. This would be a prerequisite for any evolutionary analysis and was conducted as part of the current body of work.

1.5 Aims of Research

Research into the functions of HBx has been mired in an ever-expanding list of interacting host cell proteins. As such there is no consensus on the exact function of this protein in the HBV lifecycle, and importantly the specific role this protein plays in the development of HBV-related HCC. *E.*

coli based expression systems lack the appropriate mechanisms for carrying out post-translational modifications such as disulphide bond formation, glycosylation, and phosphorylation to name a few. Therefore eukaryote based expression systems were chosen so that expressed proteins would closely resemble HBx found in infected hepatocytes. Concurrent to this, bioinformatic and evolutionary analysis of HBx was undertaken to understand HBx proteins more closely.

Previous attempts at expressing HBx in *E. coli* suggested that the expression and purification of HBx using eukaryotic cells would be technically very challenging due to the biochemical nature of HBx. However it was felt that progress could only be made on the expression, function and structure of HBx through the use of technically more difficult methodologies. The results from this work significantly improved the understanding of the nature of HBx protein, provided new avenues for the production and purification of eukaryotic HBx, and uncovered vital considerations for the design and execution of HBx studies. This thesis included the following aims:

1. Exploration into the suitability of *S. cerevisiae* and mammalian cells for the large-scale expression and purification of HBx.

2. An analysis of the HBx amino acid sequence to determine whether there were any predictors which could help explain the technical difficulties in working with this protein. This included protease sensitivity and stability.

3. A computational evaluation of regions of structural disorder within the HBx protein. This could supplement any findings of protease sensitivity, but also provide information regarding the nature of interactions between HBx and cellular targets.

4. The use of solubility-enhancing tags in combination with mammalian Semliki Forest Virus expression systems was evaluated for their combined suitability in producing large quantities of HBx in eukaryotic cells.

5. A study of the recent evolution of HBx was conducted to determine what types of substitutions are prevalent in HBx, and how these relate to the overlapping reading frames of *polymerase* and *pre-core*. Additionally, the presence of adaptive and negative evolution was assessed at sites within HBx and along the length of HBx.

6. Determination of whether there is independent evolution in HBx. This would be suggestive of differences in evolutionary pressures between the genotypes of HBV.

Chapter 2

Exposition on the Expression, Purification and Detection of Recombinant HBx

2.1 Introduction

Despite 30 years since the discovery of HBx, elucidation of potential functions of this protein remain controversial (reviewed in[59,77]). Major discrepancies in HBx related literature are common (reviewed in[78]), and are largely due to poor availability in methodologies e.g. the lack of HBV infectable cell lines, and reliance on overexpression systems to identify HBx function. This has created a fabric of data that is not biologically relevant and is riddled with artefactual findings. Some of the data are discussed briefly below to illustrate the gaping need for undertaking technically difficult experimentation to uncover the physiological functions of HBx.

1) Discrepancies arose in attempts to describe the presence of enhancer elements in HBV. Enhancer elements are typically required to increase basal levels of transcription by coordinating DNA-protein interactions in *cis* or in *trans* on upstream and downstream ORFs. The enhancer 1 (EnhI) element of HBV lies upstream of the X ORF and is itself overlapped by several cis-elements including enhancer 2 (Enh II)[79]. Di *et al*[80] were unable to identify a functional enhancer element in WHV homologous to EnhI in HBV, but they did confirm the presence of an enhancer analogous to EnhII. Opposing data was generated by Murakami and colleagues who found the predicted enhancer sequences were able to form DNA-protein complexes of several synthetic DNAs[81] suggesting that the predicted enhancer elements were functional. It is not clear why these differences were observed.

2) Sub-cellular localisation of HBx also remains controversial, with separate studies showing perinuclear localisation[82], nuclear localisation[83], and a combination of cytoplasmic and nuclear localisation by Doria *et al*[84] and separately by Nomura *et al*[85]. Specifically, Nomura showed that

FLAG-tagged HBx was found in either the cytoplasm, the nucleus or both cellular compartments. These discordant results were shown more recently to be a result of the strength of the promoter used in HBx expression localisation experiments. At low levels, HBx is largely nuclear whereas at higher expression levels it shows largely cytoplasmic localisation([86,87]).

3) A variety of data on the effect of HBx on cell death and apoptosis has been reported by several groups. These reports range from the prevention of p53-dependent apoptosis by the microinjection of HBx into primary fibroblasts[88], to the induction of p53-dependent apoptosis in NIH3T3 cells transiently expressing HBx after DNA damage with low dosages of topoisomerase II inhibitor etoposide[89]. Indeed a third finding by Su *et al*[90] did not observe apoptosis in Chang liver cells expressing HBx, but found the cells were instead sensitised to programmed cell death by low levels of TNFα.

Taken together these data illustrate the inconsistencies common to HBx related "functions" and highlight the wide range of cell culture models and methodologies that have contributed to the poor concordance between these datasets. Despite these inconsistencies, HBx is important for viral replication in woodchucks[91,92] and mammalian cells[55,93,94] and critically, it has been implicated in the progression of liver carcinogenesis in individuals chronically infected with HBV. As a result of the importance of HBx in the HBV lifecycle and potential involvement in the health outcomes of chronically infected individuals, it is vital that experimentally difficult aspects of HBx are researched so that its true mechanisms of action, function and structure are identified.

The importance of isolating preparative quantities of HBx for structural determination has been hampered by the solubility and instability problems that HBx exhibits. To date, HBx has eluded high resolution X-ray crystallography and nuclear magnetic resonance studies[59] and as a result researchers know very little about the structure of HBx. Attempts to express this protein in *E. coli* have until recently led to 99% of the expressed protein being deposited in inclusion bodies[95,96], even with the addition of solubility tags such as Glutathione-S-transferase[97]. Clearly alternative systems that do not involve bacterial or insect expression[65,98] are required to generate HBx protein and this forms the basis of the results presented in the current chapter. Specifically we wanted to express HBx in eukaryotic cells to ensure correct post-translational modifications and importantly, improved solubility. In this chapter, initial efforts were directed towards expressing HBx in different *S. cerevisiae* vectors and different mammalian expression vectors. *S. cerevisiae* has to our knowledge only been used with HBx in yeast two-hybrid experiments[99,100]. *S. cerevisiae* offers distinct advantages over prokaryotic cells including eukaryotic post-translational modifications and the relative ease of manipulation and cost of culturing conditions compared to mammalian systems. Different mammalian expression vectors were evaluated for the ability to express HIS tagged intracellular HBx, codon optimised HBx and GFP fusion constructs. We intended for these expression systems to produce quantities of HBx that would allow future work to focus on the structural analysis of HBx.

2.2 Materials and Methods

Each of the constructs used in this work is described in detail below, but an overview is described here. Initially, the *E. coli* pET-ELPI (Elastin like protein and Intein) was used as the ELP permitted rapid isolation of HBx fusion protein. The eukaryotic plasmids were then constructed, with the HBx sequence being inserted into the *S. cerevisiae* pYES plasmid, with and without a 6xHIS tag. This was repeated for the *S. cerevisiae* YEp vectors. The mammalian pCI-Neo vector followed, with one plasmid containing untagged wildtype HBx, and another a codon-optimised 6xHIS tagged HBx. Lastly, a GFP-Intein-HBx fusion sequence was sub-cloned into each of the vectors mentioned, including the mammalian pCEP4 vector.

2.2.1 Bacterial Expression Vector Construction

2.2.1.1 pET Vectors

pET-ELP-I-HBx HBx was sub-cloned into the pET-ELP-Intein system developed by Banki *et al.*[101]. The elastin-like-protein (ELP) possesses reversible solubility characteristics that are determined by temperature, salt and protein concentration. Above a certain transition temperature (T_t) and salt concentration the fusion protein becomes insoluble and hence precipitates out of solution whereas below T_t and with reduced salt concentration, the protein resolubilises. This allows the rapid partitioning of the fusion protein from contaminating cellular proteins. In addition, the Intein portion of the fusion allows the target protein, i.e. HBx, to be cleaved from the ELP-Intein fusion without the requirements for additional proteases such as Factor Xa. The effect of this system on expression and HBx solubility had not been determined and so its suitability for HBx expression was further evaluated here.

A two step approach was used to generate the vectors with oligonucleotide primers purchased from Integrated DNA Technologies (IA, USA). First, HBx was PCR amplified with a forward primer containing a *Bsp1407*I site (5′ -GATC TGTACA CAACAT GGCTGC TAGGTT GTACT-3′, the *Bsp1407*I site is underlined) and a reverse primer (5′-GATCCC TAGGTT AGGCAG AGGTGA AAAA-3′) with pCR-HBVA1 1.3x as template[102]. The PCR fragment was cloned into pTZ57R/T (Appendix B.3.1) and sequenced using standard M13 primers.

Positive clones were bulk prepared as described in Appendix B.2 and HBx fragments were excised from pTZ using *Bsp1407*I and *Hind*III restriction enzymes incubated in Tango buffer (Fermentas, WI, USA) at 37 °C. The fragments were ligated in the pET-ELP-I vector that had been cut with the same enzymes (Appendix B.3).

2.2.2 Codon Optimisation

To test whether codon optimisation of HBx would make a difference to protein expression levels, we purchased codon optimised HBx sequences from GeneArt® (Regensburg, Germany). The algorithm used to optimise the sequence minimised the occurrence of cryptic poly-A sites and reduced any calculated mRNA secondary structure that could inhibit translation. The occurrence of codons with the HBx sequence was altered to reflect similar codon useage of human proteins showing high expression levels. The same modifications were made for a HBx sequence optimised for expression within *S cerevisiae*.

2.2.3 Yeast Expression Vector Construction

2.2.3.1 pYES2 Expression Vectors

pYES2 (a gift from Professor Koos Albertyn, University of the Free State, Bloemfontein, South Africa) was the initial choice in attempts to express HBx. Protein expression in this vector is strongly controlled by the GAL1 promoter, which is repressed in the presence of glucose, but induced in the presence of galactose. This fine control of expression is ideal for potentially cytotoxic proteins as the cells are allowed to reach a logarithmic stage of growth before being induced. A Kozak translation initiation sequence was inserted before each gene to ensure the correct initiation of translation[103].

Briefly, HBx was PCR amplified from a replication competent plasmid vector of the A1 subgenotype of HBV that had been constructed in our laboratory. The HBV sequence of this vector starts upstream of the HBV basic core promoter and ends downstream of the HBV polyadenylation site[102] thus encompassing 1.3-times-the-length of the HBV genome (Fig 1.4). A forward primer (5′-GATC AAGCTT ACCACC ATGGCT GCTAGG TTGTACT-3′, *Hind*III underlined) and reverse primer (5′-GATC GAATTC *AGATCT* AGAGAT GATTAG GCAGAGG-3′, *Eco*RI underlined, *Bgl*II italicised) were used to PCR amplify HBx using standard conditions (see 2.2.5). The PCR product was cloned into pTZ57R/T (Appendix B.3.1) and positive clones were verified by sequencing. The positive pTZ clone and the pYES2 backbone (Invitrogen, CA, USA) were digested with *Eco*RI and *Hind*III in Buffer Red (Fermentas, WI, USA), and the HBx fragment was ligated into the pYES2 backbone (Appendix B.3).

2.2.3.2 YEp Expression Vectors

The YEpHF secretory vector (a gift from Professor Sally Twining of Medical College of Wisconsin, USA) was chosen as an alternative to the pYES2 vector because protein expression is driven by a yeast ADH2 promoter in the former. This promoter is regulated by glucose repression and therefore requires no inducers. Additionally the presence of the α-factor leader sequence would hopefully allow HBx secretion from the cell, and hence purification from the growth medium, resulting in a rapid and cleaner preparation of recombinant HBx.

The HBx gene was PCR amplified from pCR-HBVA1 1.3x[102] using a forward primer (5′-GATC GGTACC TTTGGA TAAAAGA *CACCACCACCACCACCAC* ATGGCT GCTAGG TTGTACT-3′, *Kpn*I underlined, 6xHIS is in italics) and reverse primer (5′-ATTACT CGAGAG GCCTAT TAGGCA GAG-GTG AAAAA-3′). The PCR product was cloned into pTZ57R/T (Appendix B.3.1) and positive clones were verified by sequencing. The pTZ clone and the YEp backbone were digested with *Kpn*I and *Bam*HI in *Bam*HI Buffer (Fermentas, WI, USA), gel purified (Appendix B.2.4), and ligated using standard cloning procedures (Appendix B.3).

YEp-GFP-I-HBx To visualise the presence of HBx within induced *S. cerevisiae* cells, a fusion protein consisting of yeast codon optimised GFP ([104] purchased from EUROSCARF), the intein fragment from the pET-ELP-I vectors (section 2.2.1.1) and 6xHIS tagged HBx from pCR-HBVA1 1.3x[102] was constructed. Initially, the YEp vector was altered by removing 6xHIS and FLAG tag by PCR amplifying the 188 bp region upstream of the HIS-FLAG tag and α mating factor with the forward primer (5′-GTATA GCATGC CTATCA CATATA AATAG-3′) containing an *Sph*I site and reverse primer (5′-G GAATTC TCGAGC TTGGTG TATTAC GA-3′) containing an *EcoR*I site. This PCR fragment was cloned into pTZ (Appendix B.3.1) and sequenced (Appendix B.2.7). Both the YepHF and pTZ plasmid were digested with *Sph*I and *EcoR*I to remove the intervening 483 bp sequence from YepHF and replaced with the new 188 bp sequence from pTZ.

The GFP-Intein-HBx fusion protein was cloned into the new YEp vector using a two step process (Appendix B.3). Codon-optimised GFP was PCR amplified using forward primer (5′-AG GGATCC *ACCACC* ATGTCT AAAGGT GAAGAA TT-3′) with *BamH*I (underlined) and a Kozak sequence (italicised)[103] before the primary ATG, and a reverse primer (5′-AG GAATTC TTTGTA CAATTC ATCCAT AC-3′) containing an *EcoR*I site (underlined). The PCR amplicon was cloned into pTZ57R/T (Appendix B.3.1) and verified by sequencing. The 1024 bp Intein-HBx segment from the pET-ELP-I-HBx vector was excised using *EcoR*I and *Not*I. Concurrently, the GFP fragment from the pTZ plasmid was excised using *BamH*I and *EcoR*I and the backbone YEp plasmid was digested with *BamH*I and *Not*I. The fragments were ligated together in a 3-way ligation reaction using a 3:1 molar ratio of insert to backbone (YEp) and a 1:1 molar ratio of insert to insert as mentioned (Appendix B.3). Positive clones were confirmed using restriction digest analysis.

pYES-GFP-I-HBx The restriction sites for the generation of YEp-GFP-I- pYES2-GFP-I-HBx vector was generated as described in section 2.2.3.2.

2.2.4 Mammalian HBx Expression Vector Construction

2.2.4.1 pCI-Neo

The plasmid containing the human codon optimised HBx was PCR amplified using a forward primer containing a *BamH*I site (bold), Kozak sequence (underlined) and 6xHIS tag (italicised), 5′-AG **GGATCC** ACCACC ATGGAG *CACCACCATCACCACCAC* GCCGCC AGGCTG TACTGC-3′ and a reverse primer containing a *Spe*I site (italicised) 5′-AG *ACTAGT* GCTCAT CAGGCG GAGGTG-3′. The PCR product was cloned into pTZ57R/T (Appendix B.3.1), and positive clones were sequenced. HBx was sub-cloned out of pTZ using *EcoR*I and *Sal*I restriction enzymes in Buffer Orange (Fermentas, WI, USA) and into pCI-Neo using the same restriction sites.

2.2.4.2 pCEP4 Expression Vector

The pCEP4 vector (Invitrogen, CA, USA) was chosen as it contains the Epstein-Barr Virus replication origin and nuclear antigen (encoded by the EBNA-1 gene) which permits extrachromosomal replication in human cells. Additionally the vector also encodes the hygromycin B resistance gene for the stable selection of transfected mammalian cells.

The GFP-Intein-HBx fusion protein was cloned (Appendix B.3) into the mammalian expression cassette pCEP4 (Invitrogen, CA, USA) using a 2 step process. GFP was PCR-amplified from pCI-Neo-GFP using a forward primer (5′-AG GGATCC GTCGAC *ACCACC* ATGGTG AGCAAG GGCGAGG-3′) containing a *BamH*I site (underlined) and a Kozak sequence (italicised)[103] before the primary ATG, and a reverse primer (5′ -AG GAATTC CTTGTA CAGCTC GTCCATG-3′) containing an *EcoR*I site (underlined). The GFP PCR product was first cloned into pTZ57R/T (Appendix B.3.1) and sequenced before proceeding. The Intein-HBx fragment was sub-cloned from pET-ELPI-HBx (section 2.2.1.1) using *EcoR*I and *Hind*III restriction sites. The pCEP4 backbone was cleaved using *BamH*I and *Hind*III and the 3 products were ligated together at a 3:1 molar ratio of insert to backbone and 1:1 molar ratio of insert to insert (Appendix B.3). Positive clones were confirmed by restriction digest analysis.

2.2.5 PCR Conditions

All PCRs were carried out using Expand HiFi^{PLUS} PCR System reagents (Roche, Basel, Switzerland) on a Mastercycler (BioRad, CA, USA). PCR mixtures included 0.4 μM of each of the respective forward and reverse primers, 10 ng of template, 1.5 mM $MgCl_2$, 1 μL dNTP mix (10mM), 2.5 U of Expand HiFi^{PLUS} Taq polymerase and 1 x final volume Expand HiFi^{PLUS} Reaction Buffer (proprietary information) in a final volume of 25 μL.

PCR conditions comprised an initial denaturation step of 3 minutes at 95 °C followed by 30 amplification cycles, where each cycle had a denaturation step at 95 °C for 30 seconds, 58 °C annealing

step for 30 seconds and a 72 °C elongation step of 1 minute per kilobase, followed by a 10 minute final elongation at 72 °C.

2.2.6 Protein Induction, Purification and Detection

2.2.6.1 Expression of HBx in *E. coli* pET15b and pET-ELP Vectors

pET15B-HBx This vector was previously constructed in our laboratory, and all protocols used to express and purify HBx protein from this vector were performed as described by Capovilla *et al*[105].

pET-ELP-Intein-HBx HBx expressed from the pET-ELP-Intein-HBx construct was purified as described in[106,107] and as illustrated in Figure 2.1. Briefly, cells were pelleted by centrifugation at 10 000 x g and 4 °C, resuspended at 10 mL per 1 g wet cell paste in 10 mM Tris pH 8.5, 2 mM EDTA and 0.1 mg/mL lysozyme and frozen at −20 °C overnight. The following day the pellets were thawed and subjected to 4 rounds of sonication, each round being 30 seconds at 75 % power. The solution was centrifuged at 37 500 x g for 30 minutes to clarify the lysate. An equal volume of 0.8M $(NH_4)_2SO_4$ was added and the sample was incubated in a waterbath at 37 °C for 10 minutes. The sample was centrifuged at 10 000 x g for 10 minutes. The pellet was resuspended in cold buffer, and the salting out procedure was repeated until the pellet resuspended without precipitates clouding the sample. To encourage cleavage of HBx from the intein fusion, the pellet was resuspended in cold PBS cleavage supplemented with 40 mM Bis-Tris pH 6.2, 2 mM EDTA and incubated at room temperature overnight. The following day the ELP-Intein protein was salted out as described above and the supernatant was dialysed to remove the $(NH_4)_2SO_4$. HBx was applied to a G-25 Sephadex column (1.0 cm diameter, 8cm length) as a final cleanup and buffer exchange into the original lysis buffer. Fraction samples containing protein were quantified using a BCA assay, concentrated using a 5 kDa MWCO Amicon ultrafiltrator to at least 1 mg/mL, and snap frozen for storage at −70 °C.

2.2.6.2 Expression of HBx in *S. cerevisiae* pYES2 and YEp Vectors

For pYES2 vectors, discrete colonies of transformed (Appendix B.1.4) YVH10 *S. cerevisiae* were picked from SCM-URA selective plates, inoculated into SCM-URA liquid media and grown at 300 rpm at 30 °C until an OD_{600} near to 0.8 was reached. An aliquot of this culture was inoculated into YP media plus 2% galactose to a final OD_{600} of 0.4 and incubated at 300 rpm overnight at 30 °C. Cultures were pelleted at 1500 x g for 5 minutes and 4 °C and either stored at −70 °C or resuspended in Y-PER (Pierce Chemical, Rockford, IL) and processed according to the manufacturers instructions.

For YEp vectors, discrete colonies of transformed YVH10 *S. cerevisiae* were picked off SCM-TRP selective plates and inoculated into SCM-TRP liquid media, grown at 300 rpm at 30 °C for 48 hours,

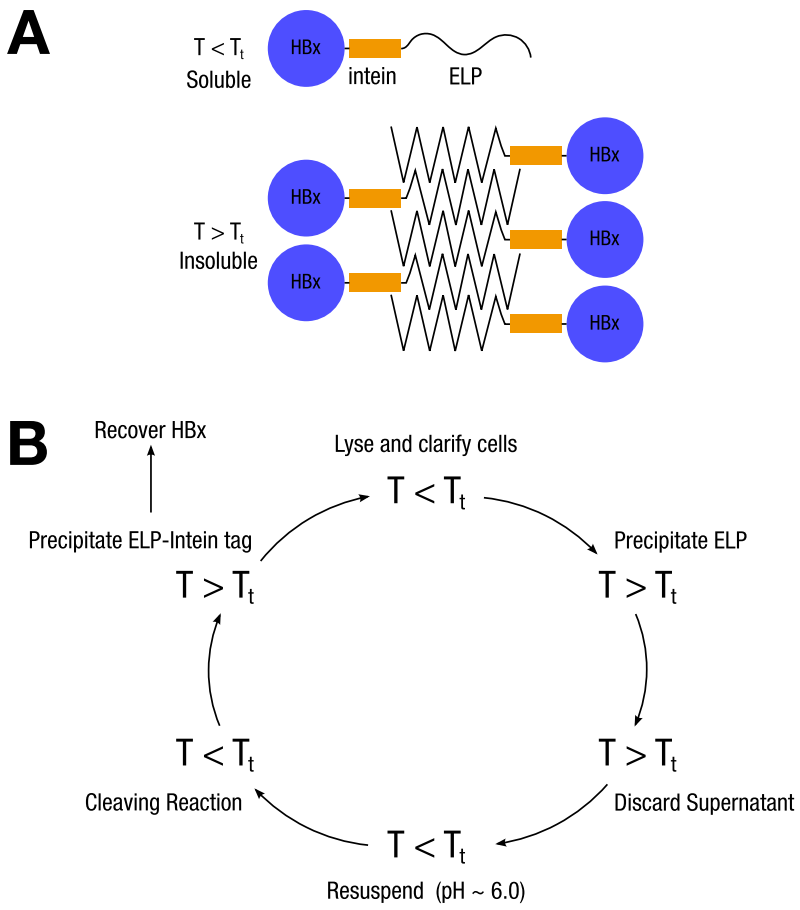

Figure 2.1: **A** - Basic description of the functioning of ELP tags adapted from[101] whereby the transition temperature (T_t) is the temperature above which the ELP fusion becomes insoluble, and below which remains is soluble. This is dependent on salt concentration e.g. $(NH_4)_2SO_4$ or NaCl, temperature and concentration of ELP protein. **B** - The solubility phase shift allows rapid separation of ELP-tagged proteins from contaminating cellular proteins, as a removal of salt and reduction in temperature causes the ELP tagged protein to resolubilise. Progress to the cleavage reaction can proceed in less than one hour under optimal conditions.

until a saturated culture was reached. A 1:20 dilution of saturated culture was diluted into fresh YPD + 3% glycerol and grown at 300 rpm, at 30 °C for 48 hours. Following induction, cultures were pelleted at 1500 x g for 5 minutes and 4 °C and either stored at −70 °C or resuspended in Y-PER (Pierce Chemical, Rockford, IL) and processed according to the manufacturer's instructions. The supernatants of the YEp vector were also stored or tested for HBx presence (see 2.2.6.4).

2.2.6.3 Expression of HBx in Mammalian Vectors

Forty eight hours after Lipofectamine™ 2000-based transfection (Appendix B.4.1), adherent HEK293 mammalian cells were resuspended using 1xPBS with 1mM EDTA. Cells were pelleted by centrifugation at 800 rpm for 3 minutes. The cell pellet was resuspended in radio immunoprecipitation assay (RIPA) buffer (50mM Tris-cl pH 7.4, 150mM NaCl, 1% NP40, 0.25% Na-deoxycholate, 1mM PMSF) and incubated at 4 °C for 5 minutes. The solution was sonicated at the lowest power setting to ensure cell lysis and to reduce viscosity. The solution was centrifuged at maximum speed and processed as needed.

2.2.6.4 Protein Concentration from Growth Medium Supernatant

Protein from supernatant was precipitated by adding one volume of 100 % trichloroacetic acid to four volumes of cell free supernatant. This was mixed by vortexing for 5 seconds and incubated for 10 minutes at 4 °C. The sample was centrifuged for 10 minutes at 16 000 x g. The pellet was washed twice with cold acetone and left to air dry. The dried protein pellet was resuspended in a solubilisation buffer (3 % SDS, 100 mM Tris, pH 11, 3 mM dithiothreitol)[108] at room temperature with gentle mixing overnight .

2.2.6.5 6xHIS tagged proteins

6xHIS tagged protein samples with volumes of $<$2 mL were purified using Ni^{2+} charged magnetic beads purchased from Genscript (NJ, USA) according to manufacturer's instructions. Larger sample volumes were batch purified with Ni^{2+} charged, cross-linked sepharose 6B (Sigma-Aldrich, MO , USA). The resin was washed once with Buffer A (20 mM Na-phosphate, pH 7.5, 2 mM β-mercaptoethanol, 10 mM Imidazole), twice with Buffer B (same as Buffer A but with 60 mM Imidazole) and finally eluted with Buffer C (same as Buffer B but with 500 mM Imidazole). The resin was washed with gentle agitation in a glass beaker and collected using vacuum filtration on a porous sintered glass funnel.

2.2.6.6 Western blotting

SDS-PAGE and Western blotting was performed as described in Appendix B.5.2 with the following modifications: a variety of blotting conditions were tested to allow detection of HBx during Western blotting; different blotting membranes were used including Nitrocellulose, 0.22 μm (used for proteins smaller than 20 kDa) and 0.45 μm PVDF; the protein component of blocking solutions was alternated between BSA and milk powder, as it is known that some antibodies react adversely to one blocking agent but not the other[109]; incubation of the primary antibody was altered between room temperature for 1 hour and overnight incubation at 4 °C, as this is known affect non-specific binding of primary antibodies[110].

2.2.6.7 Mass Spectrometry Protein Identification

As an alternative to antibody detection of HBx protein, samples for mass spectrometric analysis were prepared as follows: HEK293T cells were lipid transfected with the pCEP4-GFP-I-HBx-6xHIS using Lipofectamine™ 2000 under standard conditions (Appendix B.4.1). After 48 hours, the cells were gently washed with cold PBS, resuspended, and placed into a 15 mL centrifuge tube. Following centrifugation at 129 x g for 3 minutes, the cells were resuspended in cold RIPA buffer. Cells were aspirated multiple times to shear DNA, and were subjected to the lowest sonication possible. The lysed cells were centrifuged at maximum 16 000 x g for 5 minutes, and the supernatant was transferred to a new tube for purification as described in section 2.2.6.5.

Following purification, the entire eluted sample was loaded and run on a freshly prepared 5 - 20% gradient SDS-PAGE gel. Electrophoresis was terminated when the 15 kDa MW marker was 2 cm from the bottom of the gel. The gel was then processed for aldehyde-free silver-ammonia staining[111]. Bands of interest were excised from the gel and processed for mass spectrometry using the Trypsin Profile IGD Kit (Sigma-Aldrich, MO , USA). Prepared gel band samples were processed using oMALDI-QTOF facilities at the Council for Scientific and Industrial Research, Pretoria, South Africa.

2.2.7 Immunofluorescence Microscopy Imaging

2.2.7.1 Immunofluorescence Sample Preparation

To visually detect HBx expression in mammalian cells that had been transiently transfected with the CMV promoter-containing expression constructs described above, immunofluorescence slides where set up as follows: briefly, HEK293T cells were seeded at 250 000 cells per well in a 12 well plate containing sterile glass coverslips (12mm diameter). Twenty four hours later, each well was transfected with 2 μg of HBx-expressing plasmid and 15 ng of pCI-eGFP (used to determine transfection efficiency). Forty eight hours later the media was gently aspirated, and the cells were gently

washed twice with 1x PBS, fixed for 30 minutes in freshly prepared, filtered, 4 % paraformaldehyde in PBS. Cells were washed twice in 250 μL of 1x PBS for 5 minutes each. Fixed slides were stored at 4 °C. *S. cerevisiae* cells were prepared according to Atkin[112].

2.2.7.2 Antibody Staining of Fixed Cells

The development of multiple monoclonal antibodies that recognise different antigenic sites on HBx has proved difficult as the protein has notoriously low immunogenicity[78]. As a consequence, substantial efforts have been applied in our laboratory to develop hybridomas which produce monoclonal antibodies recognising different antigenic sites on HBx. Currently, we have access to four different HBx antibodies, three of which bind unique antigenic sites (S1, S2 and S3), and a fourth which partially overlaps with the previous three (6.1). The HBx antibodies had been partially purified using 6xHIS tagged, urea denatured, *E. coli* HBx, attached to Ni^{2+} affinity resin. They were resuspended at 0.5 mg/mL in PBS and stored at $-20\,°C$ (see section2.2.6.5). A working stock of these antibodies was prepared at a dilution of 1:100 in PBS with 2% BSA. One anti-HIS tag antibody (THE™ HIS Tag Antibody, Genscript, NJ, USA) was used to detect protein expressed from pCI-HHBx and GFP-Intein-HBx constructs. A 1:500 working stock of this antibody was prepared in PBS with 2% BSA. Secondary antibodies included a rabbit anti-mouse conjugated to Cy3 using a Lightning Link Kit (Innova Biosciences, Cambridge, UK) and Cy3 AffiniPure Rabbit Anti-Mouse IgG (H+L) (min X Hu Sr Prot Jackson Immunoresearch Laboratories Inc. PA, USA). A working stock of the secondary antibodies was prepared at 1:200 in PBS with 2% BSA. Three washes with PBS + 0.2% BSA, of 1 minute each were performed in between antibody incubations. The final wash was done in PBS + 1x DAPI to stain the nuclei. Cells were then mounted with FluorSave™ (Merck, Damrastadt, Germany) and stored at $-20\,°C$ until processed using the microscopy facilities at the Gene Expression and Biophysics Group, CSIR Biosciences, Pretoria South Africa.

2.2.8 HBx mRNA detection using RT-PCR

Total mRNA was extracted as described in Appendix B.2.5 and B.2.6 and precipitated as follows: mRNA stored in formamide was precipitated by reducing the formamide to 30 % with RNase free dH_2O, adding 1:10 v/v of 3M sodium acetate and 2.5 volumes of 100 % ethanol[113]. Samples were placed on ice for 30 minutes, after which they were centrifuged for 30 minutes at 16 000 x g. Pellets were washed with 70 % ethanol and left to air dry. Samples were resuspended in dH_2O and adjusted to 1 μg/μl.

Reverse transcription was performed as follows: a mixture of 1 μg mRNA, 2 μl of 10 μM oligo-dT primers, 4 μl of 10 mM dNTP mix made up to 16 μl with RNase free dH_2O was heated to 70 °C for 5 minutes. The sample was briefly centrifuged, and placed on ice. To each tube, a mixture

of the following was added: 2 μl of 10x RT buffer, 0.5 μl RNase inhibitor, 1 μl M-MuLV reverse transcriptase (New England Biolabs® Inc, MA, USA). The samples were then incubated at 42 °C for 1 hour. Samples were inactivated by heating to 80 °C for 5 minutes, after which the sample volume was adjusted to 50 μl dH_2O. Negative control reactions contained no M-MuLV enzyme, and were otherwise identically treated.

PCR conditions are described in section 2.2.5. For each PCR, 5 μl of RT sample was used together with GAPDH forward (5′-AGGGGT CATTGA TGGCAA CAATAT CCA-3′) and reverse primers (5′-TTTACC AGAGTT AAAAGC AGCCCT GGTG-3′) for mammalian samples, and PDA1 forward (5′-GGTCAG GAGGCC ATTGCT GT-3′) and reverse primers (5′-GACCAG CAATTG GATCGT TCTTGG-3′) for yeast samples. Primers used for the amplification of HBx were the same as those used in the construction of the expression vectors.

2.2.9 Protein Disorder and Protease Susceptibility Prediction

Prediction of protein disorder was carried out by analysing the amino acid sequence of HBx obtained from pCR-HBVA1 1.3x[102] using the GLOBPLOT[114], RONN[115], IUPRED[116], PRDOS[117] predictive algorithms. Websites for these algorithms can be found on the DISPROT website[118]. Protease sensitive sites were predicted using PESTFIND[119] and D-BOX Finder[120].

2.3 Results

The results described below offer a logical account of the experiments performed to express HBx in the yeast and mammalian cells. Initial expression experiments focussed on *S. cerevisiae* expression vectors. Despite the use of varied conditions for both the expression and detection of HBx, protein was not detected using Western blot analysis. This problem could be as a result of issues in the transcription of HBx DNA or translation of resultant HBx mRNA. The presence of full length HBx mRNA was confirmed in induced cells and in all assayed vectors, thus indicating that the problem lay with the translation of HBx mRNA or subsequent detection steps, rather than the transcription of the HBx DNA. A GFP fusion protein was thus constructed so that HBx could be visualised using fluorescence microscopy and thus determine if the problem lay with translation of HBx mRNA. The positive visualisation of fluorescent, intracellular GFP-Intein-HBx fusion protein offered evidence that HBx was being translated into protein in both *S. cerevisiae* and mammalian cells and thus the detection problem was a result of either HBx purification or antibody detection. This same protein was purified using the 6xHIS tag on the C-terminus and analysed using mass spectrometry (MS). Only peptides from the GFP portion of the fusion protein were identified. This result nonetheless indicated the presence of HBx due to the use of a 6xHIS tag. The MS results suggested that the antibodies used for HBx detection where not binding their respective antigenic sites under the conditions tested in Western blots.

To determine whether the antibodies used could bind to HBx before any purification or detection steps mentioned thus far, immunofluorescence microscopy was performed on induced *S. cerevisiae* and mammalian cells using the four in-house anti-HBx antibodies and a commercial anti-HIS antibody that had been used on Western blots. Results strongly suggested that all the antibodies did indeed bind to HBx under the conditions tested.

Detailed consideration of all results indicated that the monoclonal antibodies used for Western blots did not bind eukaryotic HBx when exposed to the denaturation steps during Western blotting, whereas *E. coli* expressed HBx was not affected by those same denaturation steps. A detailed description of the results obtained in this work are presented below.

2.3.1 pET-ELPI-HBx protein purification

Although measurable quantities of HBx were purified from the pET-ELPI-HBx system (Fig 2.2-C), this vector was not used further due to reproducibility problems and significant protein losses during the salting-out steps. Specific problems included inconsistent cleavage of HBx from the ELPI backbone, loss of HBx when applied to a G-75 column and difficulties in re-solubilising the ELPI-HBx protein pellet after salting out. The ELPI system, like other tags, does not offer a purification panacea for all proteins. The behaviour of the tag has not been perfected, and any difficile biophysical properties of

the attached protein (i.e. HBx) are not nullified by the presence of the ELPI tag.

2.3.2 Evaluation of *S. cerevisiae* Expression Vectors

2.3.2.1 pYES2 vectors

As summarised in Table 2.1, numerous expression conditions and plasmid vectors were used in an attempt to express HBx in eukaryotic cells. Initially, the galactose-driven pYES2 system was used, and in spite of variations to the expression conditions such as expression temperature (25 °C to 37 °C), induction time (4-24 hours), growth media (defined and complex media), different *S. cerevisiae* protease deficient strains (YVH10, BJ3505 and CEN.PK2) and induction ODs (0.1-0.6), the presence of the HBx was never detected by Western blotting using the anti-HBx antibodies developed in-house, or the commercial anti-HIS (Genscript, NJ, USA) and anti-HBx (Acris Antibodies, Germany) antibodies. This is shown clearly in Figure 2.2-A. The non-detection was surprising considering that GAL promoters are the most tightly regulated and powerful promoters in the GAL gene family (reviewed in[121]). Typically, induction with galactose causes a 1000 fold increase in mRNA transcription[122]. The possibility of HBx being present in the insoluble fraction was also tested, but again no HBx was detected by Western blot (data not shown). Attempts to concentrate any HBx that may have been expressed were also tried using HIS purification and TCA precipitation methods, but this also did not yield any positive results possibly due to HBx instability and non-binding of primary antibodies to HBx.

2.3.2.2 YEp vectors

In an attempt to make the purification of HBx protein easier, the protein was cloned upstream and inframe of a secretory α-mating factor. The YEp vector is driven by the ADH2 promoter, which is repressed 100 fold by glucose[123], making it useful in the expression of toxic proteins, such as may be the case with HBx[124]. Similar conditions tested for the pYES2 vectors were tested for the YEp vectors including a greater range of induction times (48-96 hours). Protein samples analysed by Western blot yielded the same results as for the pYES2 vectors, except that a supernatant fraction was included to test for secretion. Protein expression was not detected by Western blot under any of the conditions tested (Fig 2.2-A). Attempts to concentrate any HBx that might have been present in the supernatant using batch HIS purification and Amicon ultrafiltration (Millipore, USA, 5kDA MWCO) methods also did not yield any detectable protein.

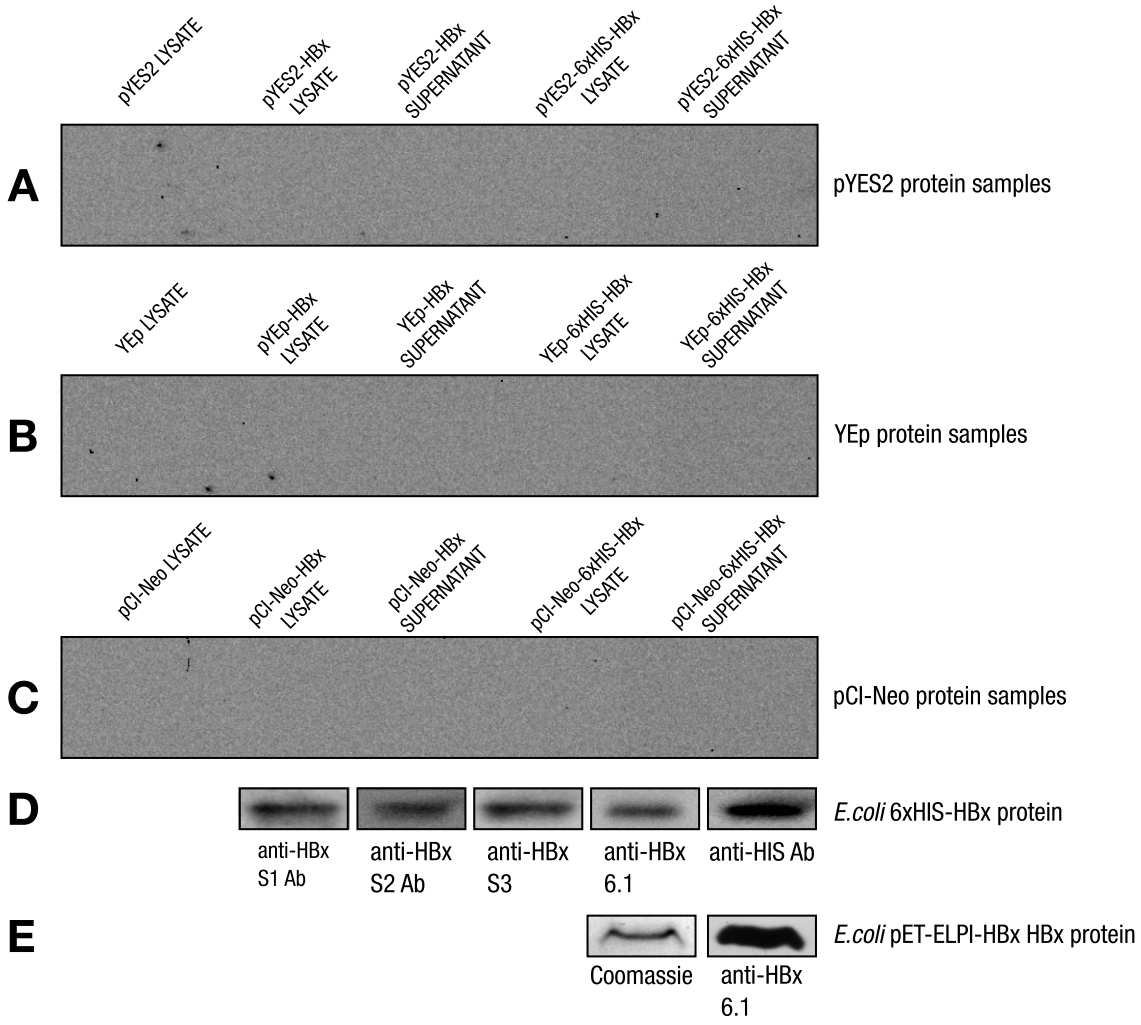

Figure 2.2: Representative compilation of the Western blots obtained using the anti-HBx antibodies. **A** - Western blot of pYES2 protein samples, including lysate and supernatant with anti-HBx 6.1 antibody. No HBx was detected in any of the samples. **B** - Western blot of YEp as in A. **C** - Western blot of pCI-Neo proteins samples as in A and B. **D** - Western blot of *E coli* 6xHIS-HBx protein samples using various anti-HBx antibodies. HBx was detected in all samples. **E** - Coomassie stain and Western blot of *E. coli* pET-ELPI-HBx protein using anti-HBx 6.1 antibody. HBx was detected using both methods.

Table 2.1: Summary of different HBx expressing clones, experimental conditions used and results obtained.

Construct	Cell line	Tag	Induction time	Media	Temperature	WB[1]	mRNA	Purification
pET15B-HBx	*E. coli* BL21	HIS	4 hours	LB	37 °C	YES	NA	Urea-HIS
pET-ELPI-HBx	*E. coli* BLR-DE3	Elastin	48 hrs	TB	25 °C	NO	NA[1]	Salt
pYES2-HBx	*S. cerevisiae* YVH10	None	12 hours	YPG[1]	30 °C	NO	YES	None
pYES2-HHBx	*S. cerevisiae* YVH10	HIS	12 hours	YPG	30 °C	NO	YES	HIS
pYES2-GFP-I-X	*S. cerevisiae* YVH10	GFP, HIS	12 hours	YPG	30 °C	NO	YES	HIS
YEp-HBx	*S. cerevisiae* YVH10	HIS	48-72 hours	YPD[1]	30 °C	NO	YES	None
YEp-HHBx	*S. cerevisiae* YVH10	HIS	48-72 hours	YPD	30 °C	NO	YES	HIS
YEp-GFP-I-HBx	*S. cerevisiae* YVH10	GFP, HIS	48-72 hours	YPD	30 °C	NO	YES	HIS
pCI-Neo-HBx	HEK293T	None	48 hours	DMEM	37 °C	NO	YES	None
pCI-Neo-HBx^{co}	HEK293T	HIS	48 hours	DMEM	37 °C	NO	YES	HIS
pCEP4-GFP-I-HBx-6xHIS	HEK293T	GFP, HIS	48 hours	DMEM	37 °C	NO	YES	HIS

[1]WB - Western blot, YPD - Yeast Extract Peptone and Glucose, YPG - Yeast Extract Peptone Galactose, TB - Terrific Broth, co - codon-optimised, NA - Not Applicable

2.3.3 Western blot Analysis

The presence of HBx in each of the expression vectors was tested under different expression conditions. Results repeatedly showed no detection of HBx protein in any of the samples taken from the eukaryotic cells (Fig 2.2 A). The positive controls of *E. coli* expressed 6xHIS-HBx, HBx protein purified from pET-ELP-I-HBx vectors (Fig 2.2 B and C) and the generic size standard (EasyWestern, Genscript, NJ, USA) were easily detected. A variety of blotting conditions (Nitrocellulose, 0.2 μm and 0.45 μm PVDF, BSA, milk powder, overnight incubation at 4 °C) yielded no detection in signal from any of the samples.

2.3.4 Detection of HBx mRNA in induced cultures and cell lines

To determine whether the problem of non-detection of HBx lay with transcription of HBx DNA or translation of HBx mRNA, the presence of full-length mRNA from the various constructs was determined (Fig 2.3). In all cases, full-length mRNA of HBx was detected, and this should have resulted in translation of the mRNA. This assay was not a quantitative measurement of mRNA, but rather a qualitative indication that HBx mRNA was present, which indirectly indicated the presence of HBx protein. The pCEP4-GFP-I-HBx vector full-length 1700 bp mRNA sample could not be amplified, although the HBx and GFP portions could be separately amplified (Fig 2.3-C).

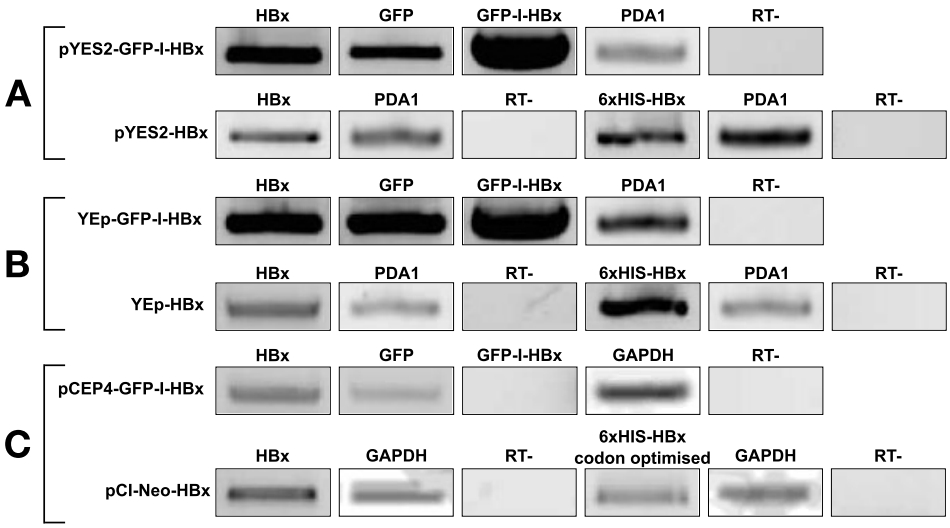

Figure 2.3: Negative exposure of RT-PCR products run on 1% agarose gels. A - pYES2 based constructs, B - YEp based contructs, and C - pCEP4 and pCI-Neo based constructs. Bands indicate the presence of full-length mRNA (Sizes of amplicons are as follows PDA1 - 672 bp, GAPDH - 100bp, HBx - 465bp, GFP - 723 bp, GFP-I-HBX - 1722 bp). RT-PCR controls included amplification of a housekeeping gene (PDA1 for *S. cerevisiae* and GAPDH for mammalian cells) plus an RT- reaction to control for plasmid DNA contamination.

2.3.4.1 GFP-Intein fusion vectors

To exclude the possibility that HBx was not being translated from mRNA or that the detection methodology was misleading, a GFP fusion protein with HBx was constructed to directly visualise expression using fluorescence microscopy. The data were promising, as the fluorescence level of GFP in all the constructs was high as shown in Figs 2.9, 2.12, 2.13. This indicated that induction was occurring and at least part of the fusion protein was being translated. However, attempts to detect the fusion protein using Western blot failed, even after protein was purified using the HIS tag on the C-terminus of the HBx segment of the fusion. This indicated either antibody binding or HBx purification problems.

2.3.5 Mass Spectrometry

To determine if the non-detection of HBx was an antibody binding problem, a purification issue, or whether the entire GFP-I-HBx fusion protein was being translated, protein that had been isolated from the mammalian pCEP4-GFP-I-HBx-6xHIS expression vector was run on an SDS-PAGE gel. After aldehyde-free silver staining, protein bands were excised, processed and analysed by MS. Of the 10 protein bands submitted, only fragments identified from the GFP section of the fusion peptide were detected (Fig 2.4). As the purification of the fusion peptide used a C-terminus 6xHIS tag, this indicated that HBx was present during the purification process and that the fusion protein was intact, but losses during cleavage of intein motif, proteolysis or autolytic cleavage during processing, resulted in an HBx concentration lower than the nanogram limit of silver staining detection. This indicated that there were antibody binding problems and/or HBx purification problems.

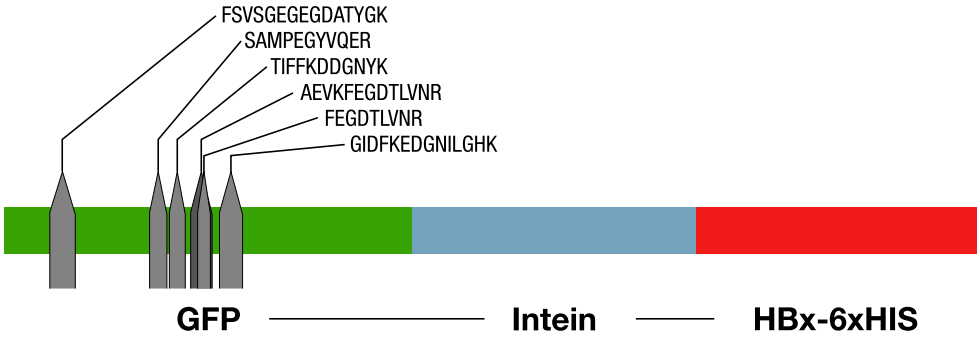

Figure 2.4: Peptides identified with greater than 90% certainty after mass spectrometry. There were no positive hits from protein bands identified from the Intein or HBx segments of the fusion protein.

2.3.6 Immunofluorescence detection of HBx protein in induced yeast cultures and mammalian cell lines

Immunofluorescence (IF) microscopy was performed to determine whether the denaturation conditions used in IF were more suited for the antibodies to bind to HBx compared to those used with Western blotting (i.e. paraformaldehyde versus SDS, heat and methanol). If IF was successful, it would show that non-fusion vectors were producing HBx protein and thus indicate that aberrant antibody binding and HBx purification problems were chiefly responsible for non-detection of HBx by Western blot.

pYES2 and YEp constructs Both the pYES2 (Fig 2.6) and YEp (Fig 2.8) vectors showed a strong signal for HBx binding with all the anti-HBx and anti-HIS antibodies tested. As expected, HBx was detected throughout the yeast cell with the pYES2 vectors, and the YEp vectors showed a strong localisation of HBx to the periphery of the cell suggesting that the α mating factor secretion mechanism was functional.

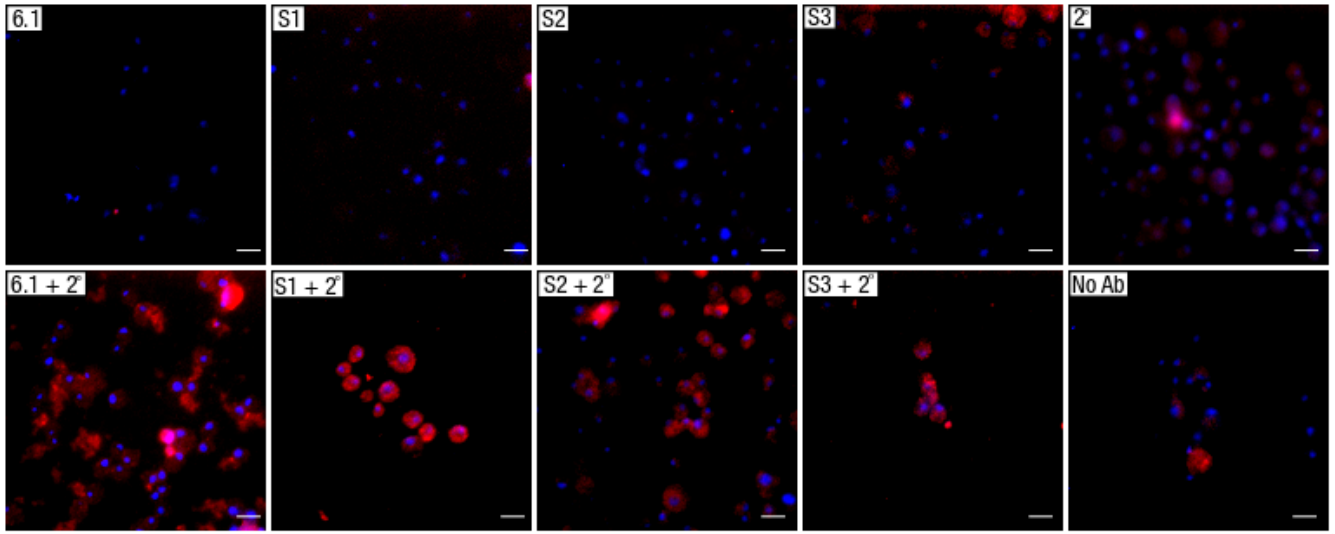

Figure 2.5: Immunofluorescence microscopy images of HBx staining in *S. cerevisiae* cells expressing the pYES2-HBx vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in samples containing primary or secondary antibody alone. Samples containing both primary and secondary antibodies showed significant Cy3 fluorescence. Magnification is 100x and scale bar = 5μm.

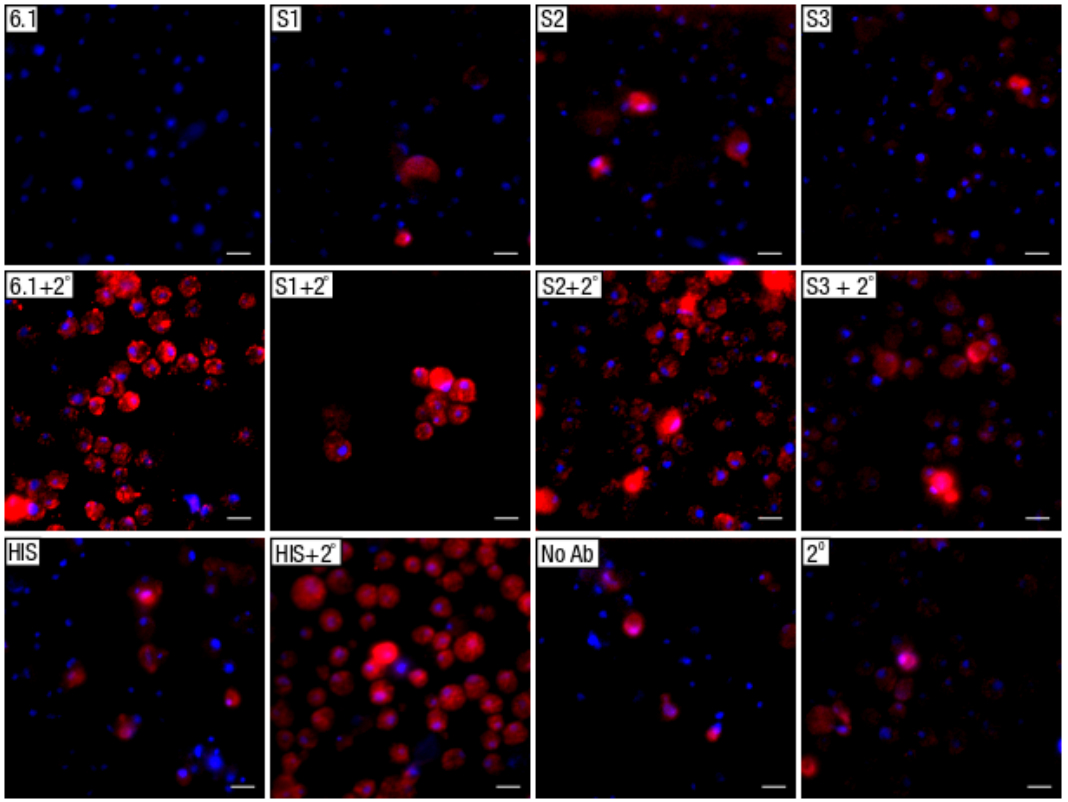

Figure 2.6: Immunofluorescence microscopy image compilation of the *S. cerevisiae* non-secreted, 6xHIS tagged, pYES2-HHBx vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. There was minimal Cy3 fluorescence in samples only containing primary or secondary antibody alone. All samples of this vector showed significant Cy3 fluorescence by the secondary antibody bound to the anti-HBx and anti-HIS antibodies. Magnification is 100x and scale bar = 5μm.

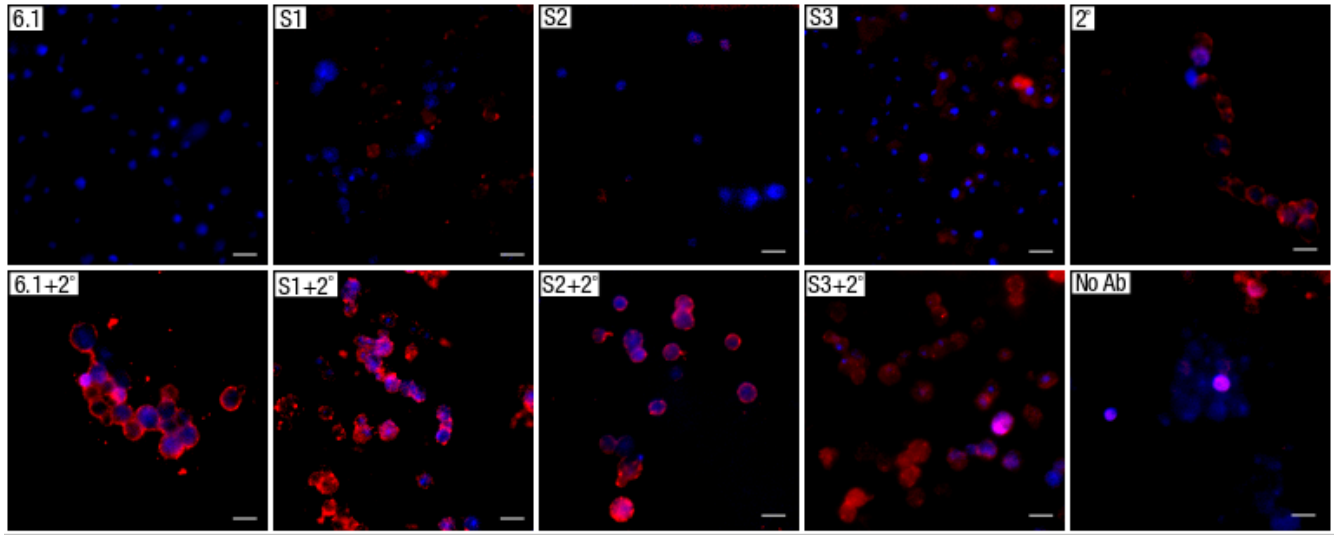

Figure 2.7: Immunofluorescence microscopy image compilation of the yeast secretory YEp-HBx vector. Antibodies specific for each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. The Cy3 signal is lower in the samples containing both primary and secondary antibodies compared to the other yeast vectors. Interestingly, the Cy3 fluorescence was localised to the cell periphery. Magnification is 100x and scale bar = 5μm.

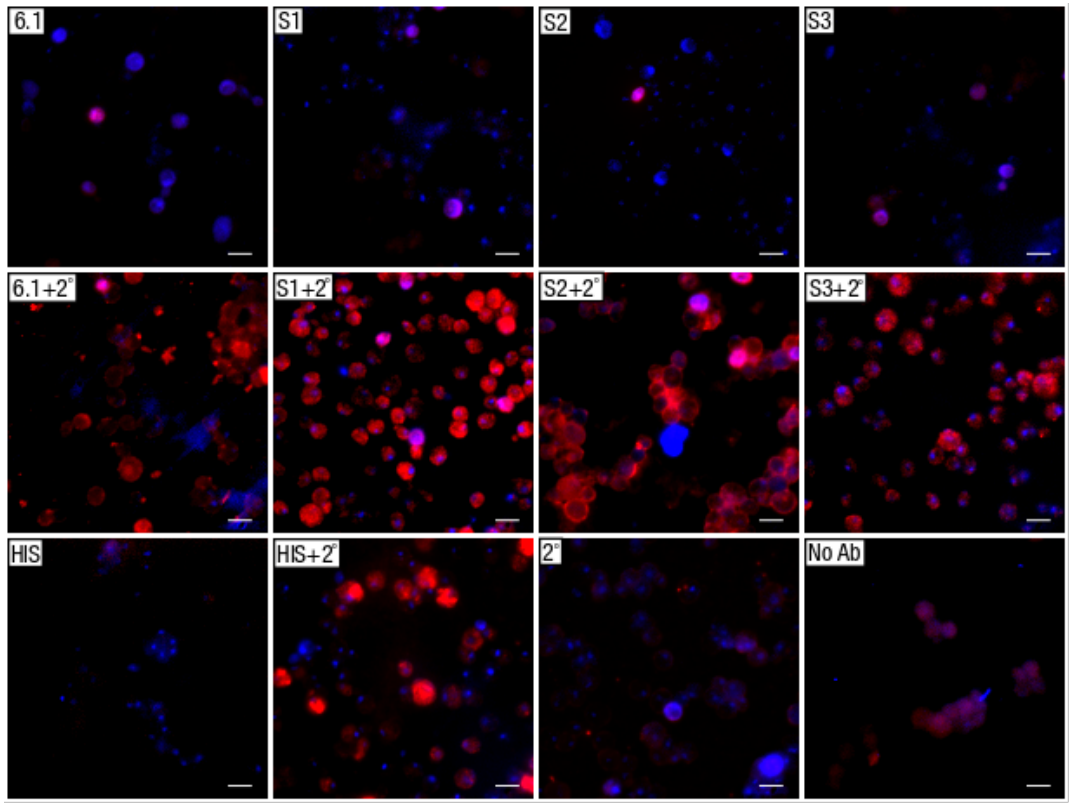

Figure 2.8: Immunofluorescence microscopy image compilation of the yeast, secreting, HIS tagged YEp-HHBx vector. Antibodies specific for each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. There is minimal Cy3 fluorescence in samples containing only primary or secondary antibody. Samples containing both antibodies show a strong Cy3 fluorescence. Cy3 was not localised to the cell periphery to the same degree as Fig 2.7, but Cy3 signal was strong with all the antibodies used. Magnification is 100x and scale bar = 5μm.

Mammalian pCI-Neo and GFP-Intein constructs Co-localisation of the GFP (green) and HBx (Cy3, red) channels of GFP-Intein constructs are shown in yellow (S2, S3 and 6.1 + Secondary in Fig 2.9). Interestingly, the anti-HBx S1 antibody did not show Cy3 fluorescence, indicating that the antibody did not bind to HBx, possibly due to site availability. In the pCI-Neo constructs, the S2, S3 and 6.1 anti-HBx antibodies for the wildtype pCI-Neo-HBx showed very weak Cy3 fluorescence with samples containing both primary and secondary antibodies (Fig 2.10).

The only antibody indicating HBx presence in Fig 2.10 was S1. Interestingly, although the codon optimised pCI-Neo-6xHIS-HBx (Fig 2.11) did not indicate Cy3 fluorescence to the levels of Fig 2.9, it nonetheless showed an increase Cy3 fluorescence over the wildtype (Fig 2.10) as indicated by fluorescence intensity.

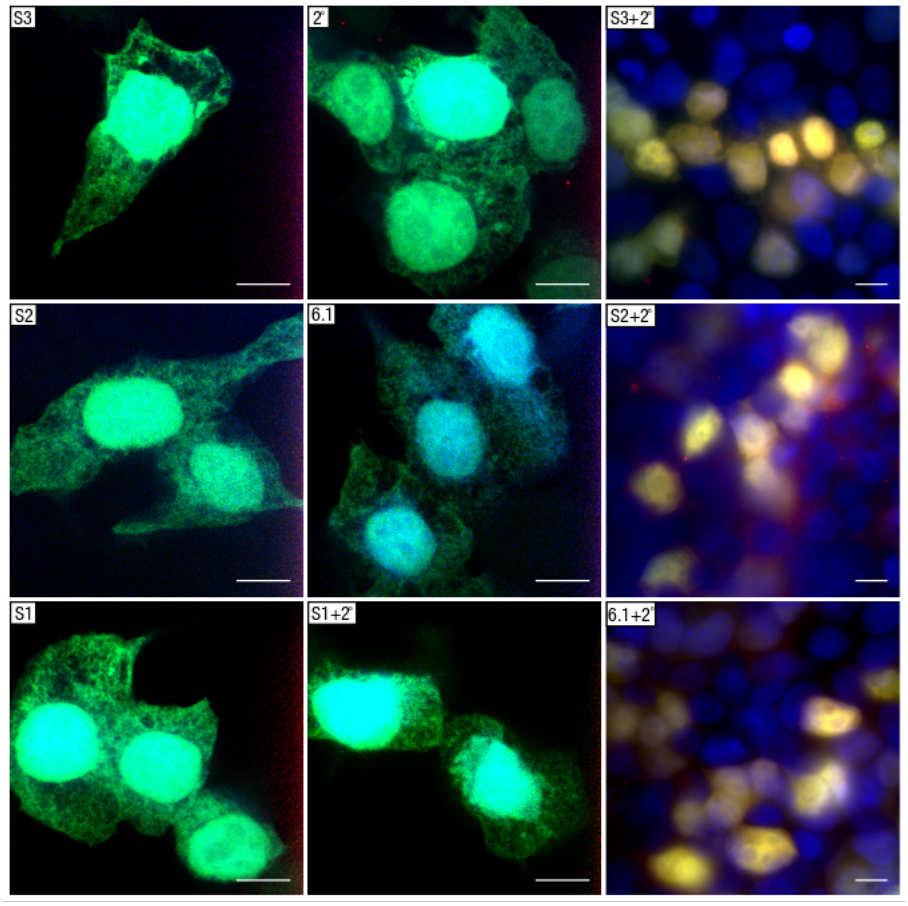

Figure 2.9: Maximum intensity projections (MIP) from Z-stacked immunofluorescence microscopy images of HEK293T cells expressing the pCEP4-GFP-I-HBx-6xHIS fusion vector. Antibodies specific to each image are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. Antibodies 6.1, S2 and S3 show good co-localisation of GFP and Cy3 fluorescence whereas S1 plus secondary shows no Cy3 fluorescence at all. Non-specific Cy3 fluorescence is low as shown by the primary, secondary and non-antibody controls. Magnification is 100x for the two left panels, and 60x for the right panel, the scale bar = 10μm.

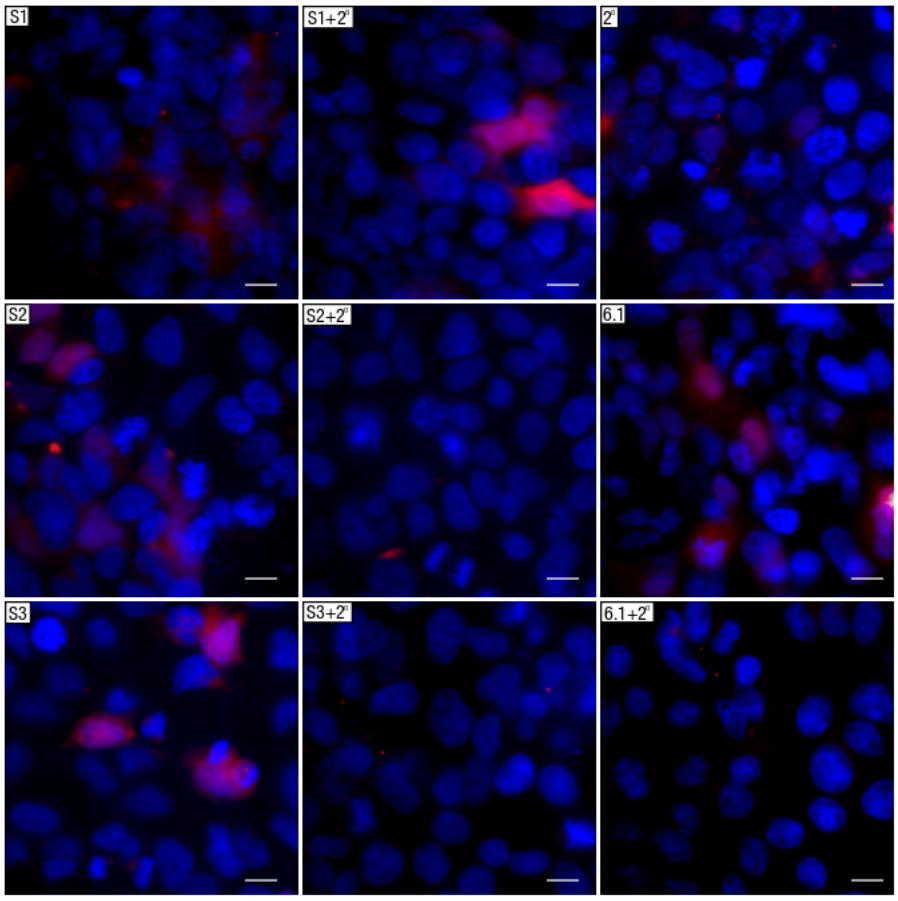

Figure 2.10: Maximum intensity projections (MIP) from Z-stacked immunofluorescence microscopy images of HEK293T cells expressing the wildtype pCI-HBx vector. Antibodies specific to each image are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence is low in the S1, S2, S3 and 6.1 plus secondary antibody samples. Only the S1 antibody shows notable Cy3 fluorescence. Magnification is 60x and scale bar = 10μm.

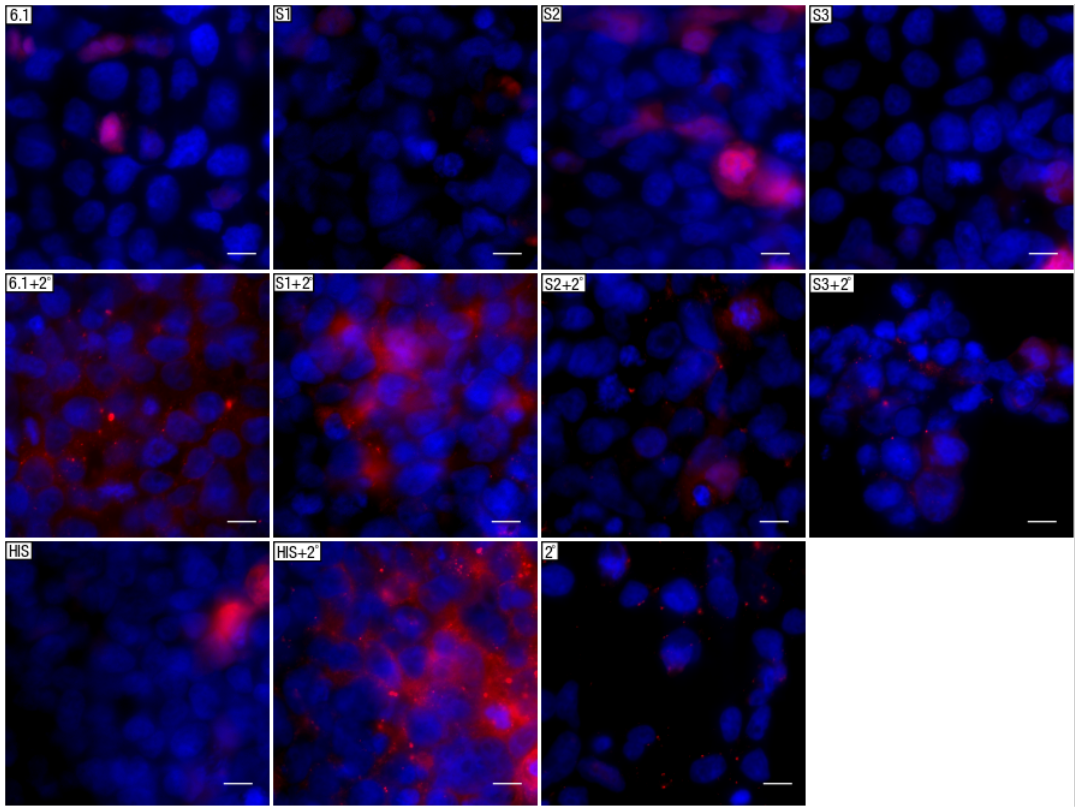

Figure 2.11: Maximum intensity projections (MIP) from Z-stacked immunofluorescence microscopy images of HEK293 cells expressing the codon optimised pCI-6xHIS-HBx. Antibodies specific to each image are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. The anti-HIS, S1 and 6.1 plus secondary antibodies showed high Cy3 fluorescence, whereas the S2 and S3 antibodies plus secondary antibody showed weak Cy3 fluorescence. Magnification is 60x and scale bar = 10μm.

Yeast GFP-Intein constructs The YEp-GFP-I-HBx vector showed a strong Cy3 fluorescence for all of the antibodies tested (Fig 2.12). The pYES2-GFP-I-HBx showed a high Cy3 fluorescence, with saturation of the 6.1 and anti-HIS antibodies. The S1, S2 and S3 antibodies showed less fluorescence intensity upon binding. Overlapping fluorescence intensities indicated co-localisation suggesting either the presence of GFP and HBx in the same region of the cell, or the presence of complete GFP-I-HBx fusion protein.

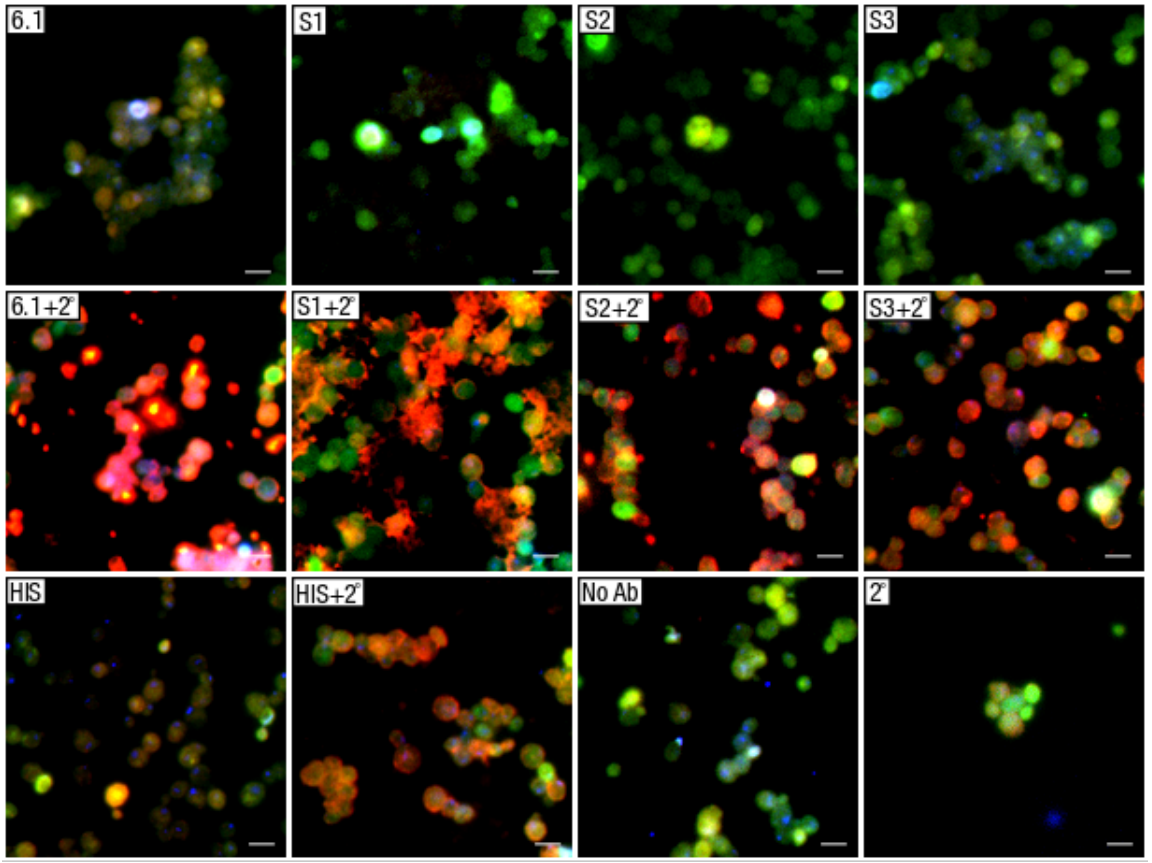

Figure 2.12: Immunofluorescence microscopy image compilation of *S. cerevisiae* expressing the YEp-GFP-I-HBx-6xHIS vector. Antibodies specific to each image are shown in the top left corner. Blue indicates DAPI staining whereas red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. A merge of the Cy3 and GFP channels indicates co-localisation of HBx and GFP (yellow/orange). Cells showed strong GFP fluorescence, and co-localisation of Cy3 in most cells containing a primary and secondary antibody. Certain cells showed only GFP or Cy3 with no co-localisation. Magnification is 100x and scale bar = 5μm.

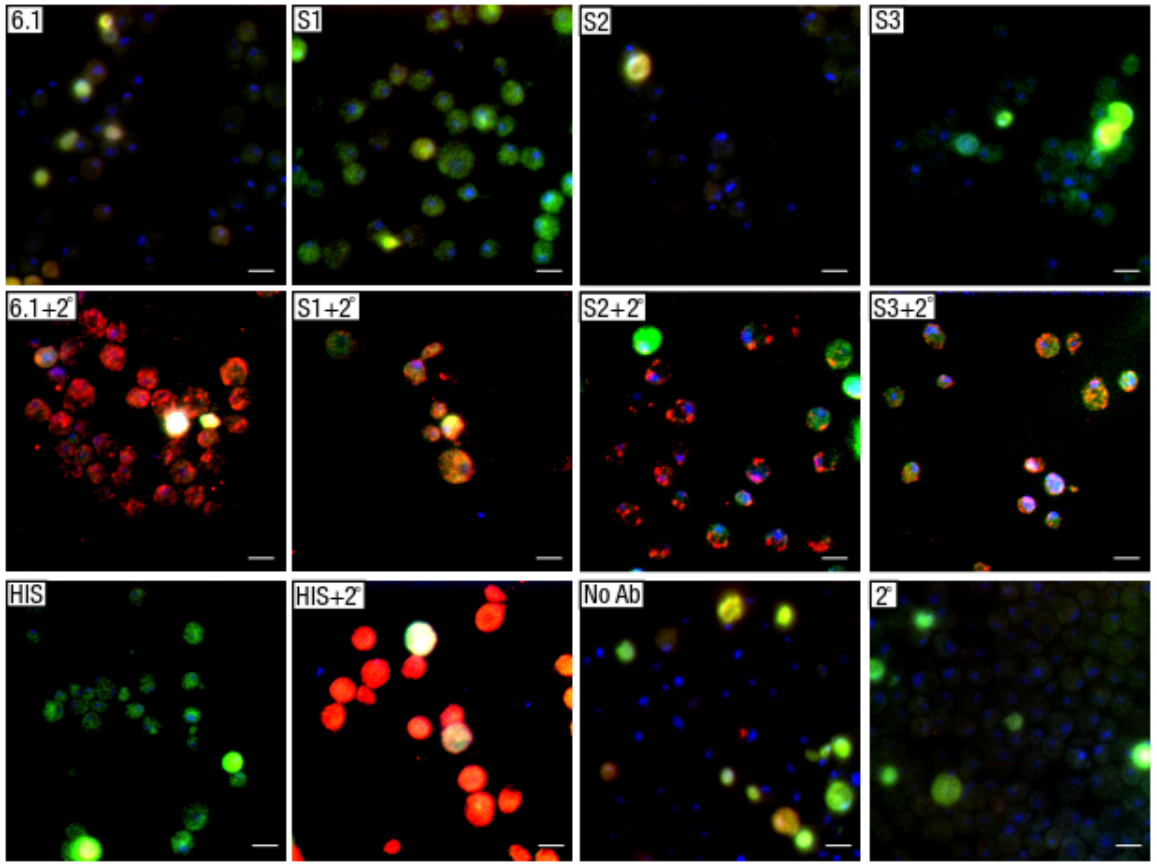

Figure 2.13: Immunofluorescence microscopy image compilation of *S. cerevisiae* cells expressing the pYES2-GFP-I-HBx-6xHIS vector. Antibodies specific to each image are shown in the top left corner. Blue indicates DAPI staining whereas red indicates anti-HBx/HIS antibodies bound to a Cy3 conjugated secondary antibody. A merge of the Cy3 and GFP channels indicates co-localisation of HBx and GFP (yellow/orange). Antibodies S1-S3 showed weak Cy3 fluorescence and co-localisation, whereas 6.1 and anti-HIS antibodies showed very strong Cy3 fluorescence. Magnification is 100x, the scale bar = 5μm.

2.3.7 Protease Sensitivity and Disorder Prediction

A variety of prediction parameters can be used to evaluate the likelihood of disordered structure of amino acid combinations in a protein. Four separate algorithms were used in the evaluation of HBx, and the results were similar with all four.

The GLOBPLOT (Fig 2.14-A) algorithm is the simplest of the four algorithms and is based on a running sum of the propensity for amino acids to be in a globular or non-globular state[114]. The algorithm predicted a disordered region (blue regions) between amino acids 25 to 52 of HBx, and a globular domain (green regions) from amino acids 58 to 145.

The RONN algorithm (Fig 2.14-B) is claimed by the authors to be the most accurate algorithm[115]. RONN (regional order neural network) compares unknown sequences with a series of sequences that have known folding states (ordered, disordered or a mixture of both). Alignment scores against these known sequences are then used to classify the disorder of each unknown sequence by using a suitably trained neural network. This algorithm predicted a disordered region between amino acids 20 and 60 of HBx.

IUPRED (Fig 2.14-C) has the underlying assumption that globular proteins are composed of amino acids which have the potential to form a large number of favourable interactions, whereas this does not occur in disordered proteins because their amino acid sequence does not allow favourable interactions to occur[116]. This algorithm shows a weaker prediction for disorder between amino acids 25 and 55 of HBx compared to the other tests.

The fourth algorithm used (Fig 2.14-D), PRDOS, utilises two predictors: the first is implemented using a support vector machine algorithm which is then combined with the second predictor which assumes the conservation of the intrinsic disorder of protein families and is an implementation of PSI-BLAST (Position-Specific Iterated BLAST). The developers of the algorithm noted that this algorithm is especially suited for the prediction of short disordered regions[117]. This algorithm predicted a disordered region between amino acids 25 and 45 of HBx.

The algorithms described above consistently predicted that amino acids between 20 and 50 represent a potentially disordered region. A prediction of HBx protease sensitivity was also performed to determine whether proteolysis could be a contributing factor in the non-detection of HBx on Western blots. HBx was computationally tested for PEST and D-BOX motifs. PEST regions are rich in proline (P), glutamic acid (E), serine (S), and threonine (T) and are associated with a short intracellular half-life[125,126]. D-BOX is a sequence recognised by the Anaphase Promoting Complex (APC/C) that in turn brings together the E2 ubiquitin-conjugating enzyme and the D-BOX. This results in protein ubiquitination and degradation. A D-BOX is characterised by amino acid sequences comprised of RLXXXXN, where R is arginine, X is any amino acid, L is Leucine, and N is asparagine.

Interestingly, analysis of HBx showed that the 20-50 bp region is predicted to have protease

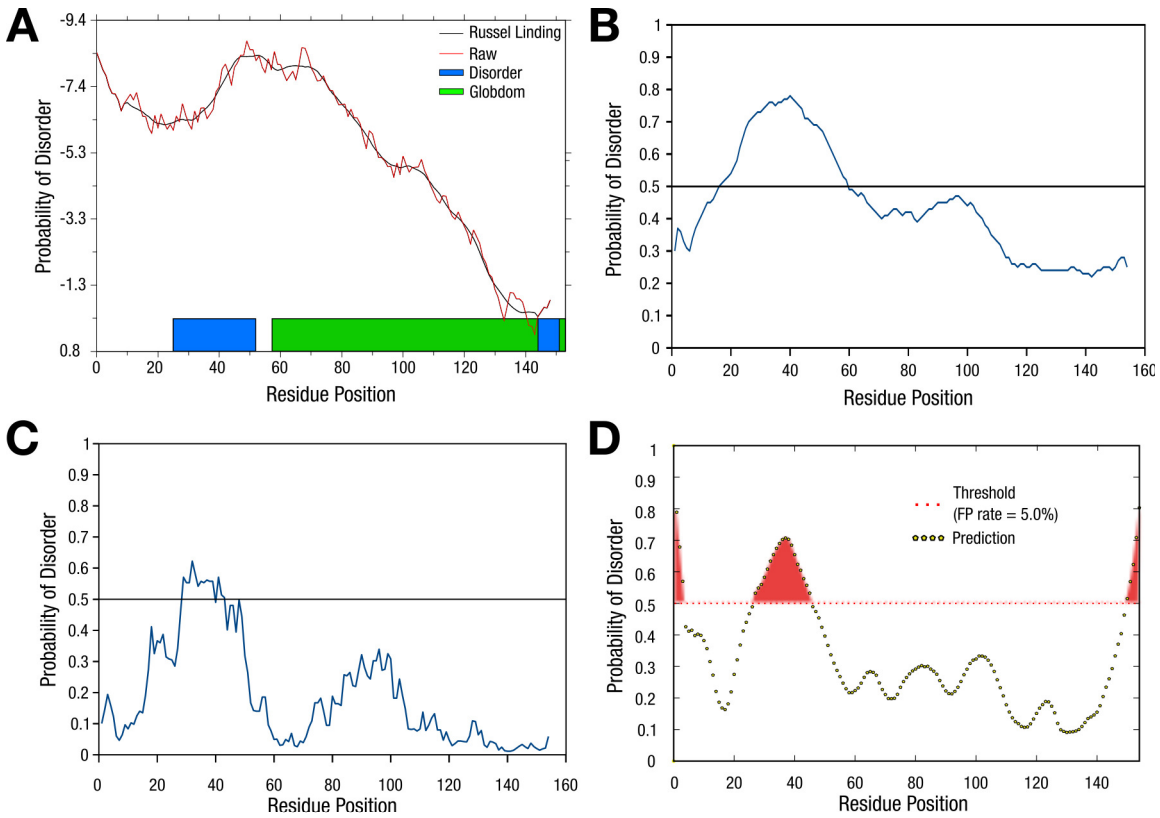

Figure 2.14: Protein sequence disorder predictions from four different algorithms **A** - GLOBPLOT. Russel-Linding (black line) plots the occurrence of coils with high temperature factors, raw data is plotted as a red line, disordered regions are shown as blue blocks, and globular domains (Globdom) are shown as green blocks, **B** - RONN - probability disorder is plotted as a blue line and the threshold for disorder is set at 0.5, **C** - IUPRED - probability disorder is shown in blue, with the threshold (black) set at 0.5. **D** - PRDOS - probability disorder is plotted as a dotted black line with disordered residues show in red above the threshold of 0.5.

sensitivity sites (PEST and D-BOX sites, Fig 2.15). Even though the scores of the predicted sites for PEST (4.17 and -12.57) and D-BOX regions were not high, it is nonetheless an indication that this region of HBx is sensitive to proteolysis. Additionally, an Instability Index (a measure of the summation of the weight values of dipeptide instability normalised over the length of the protein), shows HBx to have a measure of 59.10, classifying the protein as unstable (>40), compared to a stable protein with a value of <40 [127].

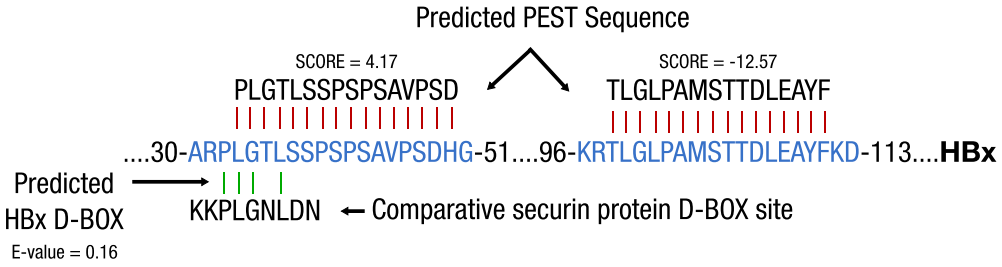

Figure 2.15: Predicted PEST and D-BOX regions in the protein HBx. HBx has two poorly predicted PEST sites (shown in red), between amino acids 33-49 (score = 4.17), and 98-111 (score = -12.57). Both sites are poorly predicted as they are below the 5.0 threshold. A D-BOX sequence (shown as vertical lines in green) is predicted to occur between amino acids 32-36, but the alignment E-value has a low score of 0.16. These sites would need to be validated experimentally to determine if they do occur.

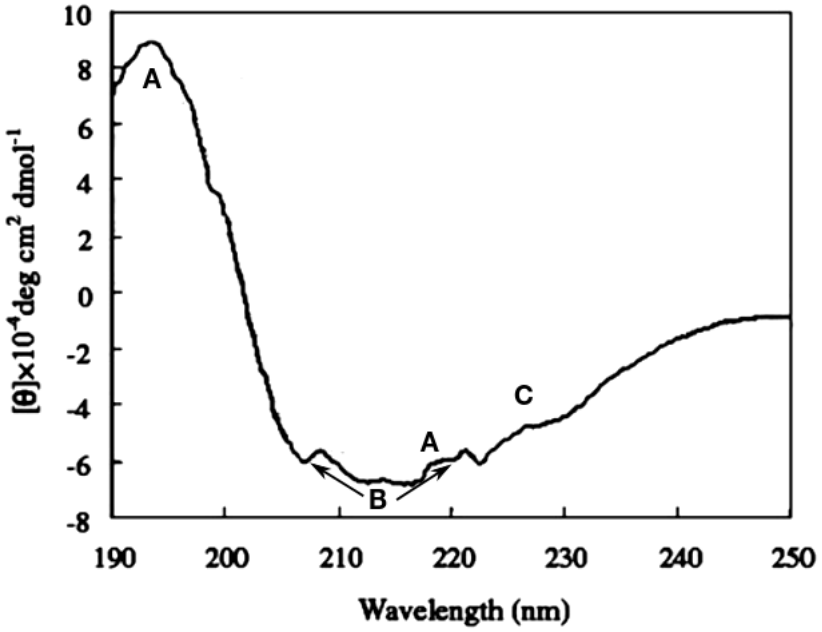

Figure 2.16: A far-UV CD-spectrograph of purified HBx from Liu *et al*[128]. The spectrograph indicates β-sheet structure(s) shown as a positive peak at 194 nm (A) and a negative shoulder at 218 nm. The two negative troughs (207 and 222nm, shown as B) are characteristic of α-helical structure(s). The weak positive peak between 220-230 nm (C) is characteristic of a disordered protein structure.

2.4 Discussion

The expression and purification of soluble and pure eukaryotic proteins is a critical point in the determination of their secondary structures. There has been little success in expressing soluble HBx protein, and there have been no published studies on the recombinant expression and purification of HBx using eukaryotic cells. HBx has unsurprisingly also evaded high resolution X-ray crystallographic and NMR structural determination. To address this in the current body of work, HBx was expressed in two different *S. cerevisiae* expression systems, with one permitting secretion of HBx. Expression was also performed in mammalian cells using either a wildtype, a codon optimised HBx or a GFP-Intein-HBx fusion protein sequence.

Despite the use of several expression systems, cell types and protease deficient *S. cerevisiae* mutants, it was a challenge to determine why neither intracellular, secreted, 6xHIS tagged nor untagged HBx protein could be detected using Western blotting (Fig 2.2-A). The use of numerous expression and detection conditions together with the positive detection of *E. coli* denatured HBx (Fig 2.2-B and C) raised concerns that transcription of HBx mRNA or the translation of HBx protein was not occurring. The premature termination of HBx transcription resulting from the presence of cryptic poly-A sites or other terminating sequence determinants were thus the first possibilities tested. RT-PCR of total RNA extractions from induced cells indicated that full length HBx mRNA was indeed present within the induced cells and thus available for translation into HBx protein (Fig 2.3). The absence of a full length mRNA for the pCEP4-GFP-I-HBx vector can be explained by amplification problems, as the individual HBx and GFP portions of the mRNA could be amplified (Fig 2.3-C). Taken together, these results suggested that aberrant transcription was not the reason for non-detection on Western blots. Whilst this experiment was not done quantitatively, it is not certain that Q-PCR would have yielded informative data as several research groups have demonstrated in both *E .coli* and *S. cerevisiae*[129] that mRNA and protein levels do not generally correlate[130,131]. As full length transcripts were detected for all of the expression plasmids, we decided to eliminate any problematic structural motifs and rare codons within the HBx sequence that could negatively affect translation. We obtained codon optimised HBx sequences for human and yeast cells. Numerous examples regarding the positive effect codon optimisation has had on recombinant protein production have been recorded for *E. coli*[132], *S. cerevisiae*[133], *Pichia pastoris*[134] and human cells[135]. Disappointingly, the benefit that codon optimisation might allow for HBx expression was nullified as the codon optimised constructs showed no improvement in detection over the wildtype HBx sequences using Western blots (Fig 2.2-A). These results suggested that the non-detection of HBx was either a result of the lack of experimental rigour, the non-translation of HBx due to undetermined factors, or problems with the methods and reagents used in the detection of HBx protein.

To eliminate the possibility of experimental rigour being at fault, two separate researchers in

different laboratories, using different reagents and equipment were tasked with expressing HBx in mammalian systems and trying to detect HBx using Western blot. In both of these instances, positive controls, which were either HBx purified from pET15B-HBx (Table 2.1) for the anti-HBx antibodies or the HIS protein ladder from Qiagen for the anti-HIS antibodies, were detected. None of the samples containing denatured eukaryotic HBx could be detected on the Western blots (data not shown, personal communication with Kubendran Naidoo and Naazneen Moolla).

The combination of (1) negative Western blots using eukaryote expressed HBx, (2) positive detection of prokaryote expressed HBx and (3) positive mRNA detection from eukaryotic vectors, raised the possibility that HBx was either not being translated from its cognate mRNA or that antibody detection was problematic. To test the first hypothesis, immunofluorescence microscopy was performed on cells transformed with the expression vectors listed in Table 2.1 using suitable negative controls to discount non-specific antibody interactions (Fig A.1 to A.6). Interestingly, the pYES2 (Figs 2.5, 2.6, 2.13), YEp (Fig 2.7, 2.8, 2.12) and pCEP4 (Fig 2.9) constructs showed a strong signal for HBx-antibody binding. This indicates quite convincingly that the translation of HBx was indeed occurring, and that the problems experienced with its detection occurred post cell lysis. To confirm these results, mass spectrometric analysis of pCEP4-GFP-Intein-HBx-6x-HIS purified protein (Fig 2.4) showed that at a minimum, the GFP portion of the protein could be detected. The non-detection of HBx using MS was unsurprising considering the previous technical difficulties in visualising HBx using silver stained SDS-PAGE gels and Western blots. As the purification of GFP depended on the presence of HBx due to the 6xHIS tag on the C-terminus of the protein, the presence of GFP could be construed as indirect confirmation of HBx presence.

These interesting results and the confounding positive HBx detection obtained from the *E. coli* pET-ELPI-HBx and pET15B-HBx plasmids provided important points for consideration. Was the native structure of eukaryotic HBx prone to aggregation and proteolysis? Alternatively, was non-detection a result of post-translational modifications to HBx such that purification and or detection attempts were thwarted? Additionally, were experimental conditions altering the antigenic sites that the anti-HBx (6.1, S1, S2 and S3) and anti-HIS antibodies recognised, thus preventing recognition? Lastly, were the screening conditions under which the anti-HBx antibodies selected inappropriate for Western blot assays of eukaryotic HBx?

To consider the first point, that HBx is possibly prone to aggregation and proteolysis, work done by Liu *et al*[128] and recently extended by Basu *et al*[136] indicated that HBx has disordered region(s) within its secondary structure, particularly between amino acids 25 to 52. This important finding was deduced from the weak positive peaks between 220-230 nm using far-UV CD-spectra of soluble HBx (Fig 2.16). It is now well established that the proportion of disordered proteins in the genomes of Eukarya range from 35-51%[137] and is thus not an uncommon occurrence. Sequence constraints (e.g. small overlapping viral genomes such as HBV) mean that having disordered re-

gions within proteins allows a greater range of interactions (e.g. protein-mediated regulation and cell control, signal transduction and transcription regulation) as compared to proteins with more rigid structures. Greater ranges of interactions are obviously incompatible with rigid and static tertiary structures[138,139]. Certainly, disorder seems to be a characteristic of many viral proteins[140] and HBx is a good example if one considers the numerous interactions ascribed to it. Intrinsically disordered or unstructured proteins (IDP/IUP) are able to adopt secondary structures depending on stimuli or interactions with a binding partner. This structural plasticity allows an exceptionally diverse range of interactions depending on the state of the cell, and is termed binding promiscuity[141] or "one to many" signalling[142].

Importantly for this work, there are technical problems for recombinant protein production involving proteins with disordered regions as they are known for their proteolytic susceptibility[143]. Even so, disordered motifs may avoid degradation through several mechanisms such as the protection of disordered sites by steric factors, the lack of protease sensitive residues within disordered sequences and the possibility that disordered regions may exist only transiently, as proteins move from one binding event to another[144]. This has been demonstrated with calmodulin which becomes sensitive to proteases when Ca^{2+} levels drop thus leaving its enzymatic partners to bind with proteins containing IQ motifs instead[145,146]. However proteasomes can degrade disordered proteins directly, without prior ubiquitination,[147–149] and an unstructured initiation site is required even for degradation of tightly folded but ubiquitinated proteins. Even though a ubiquitin tag by itself it not sufficient for proteasome degradation[150], it would be interesting to see if the co-expression of a described binding partner would reduce the chance of HBx being ubiquitinated[151] and thus degraded or whether a binding partner could prevent direct proteolysis by direct action of the proteasome. Additionally proteasome deficient mutants could be used to determine the protective effects of binding partners.

The half life of proteins is determined by a number of interacting factors not just disorder or protease sensitivity. Bioinformatic analysis of HBx weakly predicted the presence of two signal peptides that could increase HBx degradation. These include D-BOX and PEST sequences (Fig 2.15), and all predicted sites overlap with the predicted disordered region described above. D-BOX patterns tend to fall within locally disordered regions[152] and PEST sequences are known to correlate with protein disorder[153] and to be a partial determinant of protein turnover[125,126,154]. These predicted sites (amino acids 25 to 60) are within the negative regulatory domain (amino acids 1 to 50) and are next to the start of the trans-activation domain (amino acids 51 to 154)[155]. It is unclear how this disordered region is involved in either negative regulation mechanisms or if it is involved in trans-activation and/or co-activation[156] without further studies. However it is an important observation regarding the structure and function of HBx. Protein disorder may be a major determinant of protein degradation, but is by no means the only factor, even though disorder has been demonstrated to

play a bigger role than motifs like the D-BOX or PEST sequences. Protein degradation is ultimately determined by multiple factors which may carry different weighting depending on the protein[144].

From the data, it seemed however that the structural and biophysical peculiarities of HBx, whilst helping explain the inherent difficulties in working with this protein, were not an adequate explanation for its non-detection. Certainly the points discussed above may be compounding factors in the non-detection of HBx, but the shear lack of signal from Western blots suggested that the explanation had more to do with antibody detection and secondary structure modifications, rather than problems with the expression of HBx protein which would have shown a faint or degraded signal rather than no signal at all.

The issue of antibodies and antigenic sites has been in the fore recently, with projects like the Human Protein Atlas rigorously validating commercially available antibodies and finding that only 35% work under the conditions tested[157]. Although reasons for these antibody failures were not stated, it seems reasonable that careful characterisation has not taken place for many commercially available antibodies and that their detection abilities are thus overstated by companies and over-estimated by researchers. This under-characterisation of antibodies almost certainly occurs with in-house antibodies as well. The selection of antibodies is a very sensitive procedure, and care should be taken when preparing the target protein for selection. As the anti-HBx antibodies used in this project were raised against urea-denatured, *E. coli* expressed HBx, it can be reasonably assumed that antibodies that were selected would either fail to bind to the eukaryote derived HBx entirely, or would share some antigenic determinants and conformational features between the eukaryotic and prokaryotic HBx. The denatured *E. coli* HBx used for antibody selection may also have undergone further denaturation when combined with Freund's adjuvent, or undergone proteolysis when injected into mice, thus further complicating selection conditions. This could promote the selection of antibodies that recognise unfolded or degraded protein, and if subsequent assays did not test for authentic recognition of eukaryotic HBx, it is likely that antibodies were selected that recognise HBx only under certain circumstances. It follows that multiple selection methods of antibodies (i.e ELISA and/or Western blot against *E. coli* denatured and native eukaryote derived HBx) are important in determining whether antibodies work in different immunoassays. It is vital to keep in mind that studies have shown that rapidly binding monoclonal antibodies recognise epitopes on the native protein, whereas antibodies that react slowly in solution bind preferentially to the denatured form of the protein[158,159]. As experienced in this work, the anti-HBx antibodies probably recognised *E. coli* denatured HBx epitopes better than the eukaryotic native or denatured epitopes. This has been shown to occurr with other proteins such as the β_2-subunit of *E. coli* tryptophan synthase[160]. The antibodies used must have bound antigenic determinants which revealed themselves upon partial or total unfolding of denatured *E. coli* derived HBx[160], which is known to occur when a protein is immobilised on a solid phase support such as an ELISA plate. Furthermore, antigen conformation

on a solid phase may frequently be different to the conformation found in solution and that screening by ELISA, in which the antigen is attached to a solid phase, can be misleading in determining useful monoclonal antibodies[161]. Overall *et al* suggested that the presentation of epitopes to antibodies should ideally occur in solution, although if this is not possible due to protein solubility problems, then screening methods should extend to other solid supports such as Western blot[162]. Indeed indications are that this is not the first study to experience problems with antibody specificity against HBx[78]. Thus it seems that immunogenic determinants are shared between eukaryotic HBx protein and denatured *E. coli* HBx. These epitopes are easily recognised by ELISAs using denatured *E. coli* derived HBx, and immunofluorescence detection of eukaryotic HBx, but not denaturing Western blots of eukaryotic HBx. The non-detection of denatured eukaryotic HBx by Western blot seems to suggest that the problem concerns not only the antibodies used, but also the structure of the epitopes the various antibodies bind to, the antibody selection procedures used, and effects of various denaturation procedures on eukaryotic HBx structure. For clarification, during SDS-PAGE, eukaryote native HBx is heat-denatured with SDS at 95 °C, which causes a certain denatured structure, whereafter the transfer of HBx to PVDF membrane involving a 20 % methanol Towbin transfer buffer adds another level of denaturation. These successive denaturation steps could result in cumulative alterations to the HBx secondary structure, such that the antigenic sites are no longer present or recognised. This altering of epitopes of native HBx, including 6xHIS, is the most likely explanation for the non-detection of eukaryotic HBx in Western blots with the antibodies used. It is likely that the recognised epitopes of eukaryotic HBx are discontinuous (i.e. an epitope where the amino acids are in close proximity in the folded protein, but distant when unfolded). These discontinuous epitopes are most likely hidden or destroyed when certain denaturing conditions are experienced by HBx under Western blots. Overall, these results illustrate the importance of determining whether an antigen recognised by an antibody is present in both native and denatured forms of the protein[160]. Despite the technical problems regarding antibody recognition of HBx, the protein was successfully expressed in both *S. cerevisisae* and mammalian cells. Future work will need to focus on improving expression levels through additional HBx sequence optimisation and the use of other detection tags such as FLAG or c-myc that might allow detection via Western blot. This work also illustrates interventions and experiments to consider when trying to produce proteins that are difficult to work with due to solubility, stability, antigenicity, proteolytic or detection problems. Considering the stability, solubility and antigenic challenges encountered with trying to express HBx using the systems described in this chapter, work aimed at improving these was conducted and is described in Chapter 3.

Chapter 3

Expression and Purification of HBx Protein using a combination of Maltose Binding Protein Tag and Semliki Forest Virus

3.1 Introduction

The difficulty in expressing and detecting native HBx protein in eukaryotic cells detailed in Chapter 2 prompted the evaluation of alternative antibody targets for HBx detection as well as alternative systems of expression. Large fusion tags that can be detected using antibodies should ideally increase solubility and permit the rapid and simple purification of proteins such as HBx. In addition, fusion tag expression systems should be easily scaled and have been reported to show good expression for a variety of proteins.

Transient expression systems combined with solubility enhancing tags (e.g. thioredoxin, GST and NusA) allow the researcher to express and purify recombinant proteins with improved solubility, stability and rapidity. A popular solubility tag is Maltose Binding Protein (MBP) which is a native *E. coli* expressed protein that is effective in promoting the solubility and native folding of proteins fused to it[163]. MBP displays the best solubility enhancing properties compared to glutathione-S-transferase and thioredoxin when tested with a passenger group of aggregation prone proteins[163]. It is unclear whether this enhanced ability is due to chaperone-like properties[164–166] or merely passivity[167]. Nonetheless, Liu *et al* recently used this tag to express and purify soluble HBx in *E. coli*[128]. This system presented a good starting point from which to produce soluble HBx in a mammalian expression vector.

Semliki Forest Virus (SFV), is one such vector which has been used to express a wide range of

proteins, including those that are membrane bound, G-protein-coupled receptors and ligand-gated ion channels[168–170]. Importantly for HBx, SFV has been used to express the structural proteins of HBV[171], as well as the polymerase[172].

SFV-derived vectors are amongst the most studied Alphavirus vectors used for recombinant protein expression due to a number of advantages. The virus has a broad host range, and so a variety of cellular backgrounds may be used to express proteins. These vectors may be used to express protein either *in vitro* or *in vivo*. Expression typically occurs when a heterologous gene sequence replaces a viral structural protein that is required for the encapsidation of viral RNA. Current versions of the SFV vector are based on helper RNAs, and in some cases two helper RNAs[173], where the viral capsid and spike proteins have been cloned into separate vectors. Co-transfection of all the plasmids into cells results in the generation of packaged replicons which can infect cells, but cannot replicate themselves. The vectors are thus safe to work with. This approach is able to rapidly generate high titer virus stocks and the efficient replication of the viral mRNA replicon opens the possibility of infecting large numbers of cells very quickly, and thus producing substantial quantities of recombinant protein.

Considering the difficulties in detecting eukaryotic HBx described in Chapter 2, and the previous difficulties encountered in the same laboratory with prokaryotic HBx expression, we decided to utilise a solubility enhancing fusion system based on MBP. This rationale was based on the solubilisation and stabilisation of the formidable BRCA2 protein using a double MBP fusion tag[174]. In a separate study HBx was tagged with MBP[128] and expressed in *E. coli*. This differed significantly from previous work conducted in our laboratory, where prokaryotic HBx was only expressed in an insoluble form.

In this chapter, HBx was expressed as part of a MBP fusion protein in both *E. coli* and mammalian cells. The initial expression in *E. coli* was used to optimise purification of the MBP-HBx fusion protein and was based on the work by Liu *et al*[128]. This system was sub-cloned into different SFV expression constructs to determine the suitability of combining MBP and SFV to express soluble HBx in eukaryotic cells.

3.2 Materials and Methods

3.2.1 pMAL-HBx Vector Construction

The pMAL-c2 (New England Biolabs® Inc, MA, USA) vector which encodes the MBP tag and permits cytoplasmic accumulation of the fusion protein, was used to express HBx. The purification and characterisation of this procedure was recently described[128]. Briefly, to facilitate the blunt cloning of HBx into the pMAL-c2 vector, HBx was PCR amplified with a forward primer (5′-ATGGCT GCTAGG TTGTAC TGC-3′) and a reverse primer with *Sma*I (5′-CCCGGG CTATTA GGCAGA GGTGAA AAAGTT GC-3′) using pHBV-1.3x as template[102]. The PCR amplicon was cloned directly into pMAL-c2 that had been blunt digested with *Xmn*I in Buffer Tango (Fermentas, WI, USA), using a 1:5 insert to backbone molar ratio (Appendix B.3). pMAL-HBx clones with the correct orientation were sequenced using standard pMAL-c2 sequencing primers.

3.2.2 Semliki Forest Virus Expression Vectors

3.2.2.1 pSFV-b12a vector

The pSFV-b12A vector was a kind gift from Professor Cristian Smerdou (CIMA, Pamplona, Spain) and has been described elsewhere[169]. The enhancer boosts protein expression levels and is encoded by the first 34 amino acids of the SFV capsid[175]. The enhancer is fused to the amino terminal of the Foot and Mouth Disease virus (FMDV) 2A autoprotease, which permits autocatalytic cleavage of the enhancer-2A from the recombinant protein.

A schematic of the cloning procedure used to generate pSFV-b12a-MBP-HBx vector is illustrated in Figure 3.1. As the pSFV-b12a vector contains a single *Apa*I site for cloning, subcloning was initially completed in the pGEM vector (Promega, WI, USA) as follows: the secretory leader sequences contained a *Nde*I site, thus the *Nde*I site present in pGEM was deleted by incubating pGEM, linearised with *Nde*I, with high fidelity DNA polymerase (GeneAmp™ High Fidelity PCR System, Applied Biosystems, CA, USA) and 1 μL dNTP mix at 57 °C for 15 minutes. The pGEM backbone was then purified using phenol:chloroform and re-ligated (Appendix B.3). Positive clones (pGEMΔNdeI) were selected based on the inability of *Nde*I to cut the pGEM backbone.

Secretory leader sequences were included to aid in the purification of HBx from pSFV-b12A-MBP-HBx. Secretory leaders PPT, BM40 (derived from osteonectin) and CD5 were included in the cloning procedure. Initially sequences were cloned into pTZ57R/T as follows: 0.1 μM of 2 partially complementary oligonucleotides spanning the length of BM40 (5′-AG GGGCCC ATGGGC GCCTGG ATCTTT TTCCTG CTGTGC GCCGGC AGAG-3′, *Apa*I underlined) and (5′-AG GTCGAC AA *CATATG* GCCCTG GGCCAG TGCTCT GCCGGC GCACAG CAGG-3′, *Sal*I underlined, *Nde*I italicised) or CD5 (5′-AG GGGCCC ATGGGC ATGGGA AGCCTG CAGCCT CTGGCC ACACTG

TACCTG CTGG-3′, *Apa*I underlined) and (5′-AG GTCGAC AA *CATATG* GGCCAG CACGCT GGC-CAC CAGCAT GCCCAG CAGGTA CAGTGT GGCCA-3′, *Sal*I underlined and *Nde*I italicised) were extended to their full length using the GeneAmp^{TM}High Fidelity PCR System (Applied Biosystems, CA, USA) by repeating 3 amplification cycles, each at 95 °C denaturation for 30 seconds, 58 °C annealing for 30 seconds and 72 °C elongation for 10 minutes (Appendix B.3.1). Presumptive clones were sequence verified (Appendix B.2.7). The PPT leader sequence was short enough to be included in the primers used to amplify MBP-HBx for cloning into vectors.

Restriction digests of the pTZ-CD5 and pTZ-BM40 plasmids were performed using *Apa*I and *Sal*I enzymes in *Bam*HI buffer, at 37 °C overnight. The fragments were gel purified using standard procedures but supplemented with glycogen (Fermentas, WI, USA) to a final concentration of 0.05 μg/μL to aid in the precipitation of the small DNA fragments (64 bp for BM40 and 82 bp for CD5) (Appendix B.2.4). These fragments were ligated into pGEMΔNdeI and presumptive clones were confirmed by restoration of the *Nde*I restriction site and subsequent sequencing.

The MBP-HBx fusion protein was cloned into pSFV-b12A as follows: MBP-HBx was PCR amplified using a forward primer (5′-AG GGGCCC *CATATG* AAAATC GAAGAA GGTAA-3′, *Apa*I underlined, *Nde*I italicised) and a reverse primer (5′-*GTCGAC* GGGCCC CTATTA GGCAGA GGTGAA AAAG-3′, *Apa*I underlined, *Sal*I italicised). The PCR amplicon was cloned into pTZ57R/T and positive clones were sequence verified (Appendix B.3.1). MBP-HBx was excised from pTZ using *Nde*I and *Sal*I in Buffer Orange (Fermentas, WI, USA), gel purified (Appendix B.2.4), and cloned into similarly digested pGEMΔNdeI-CD5 and pGEMΔNdeI-BM40 (Appendix B.3). Positive clones from this step were digested out of pGEMΔNdeI backbone using *Apa*I and cloned into similarly digested pSFV-b12A. Positive orientation of the MBP-HBx insert was confirmed by digesting pSFV-b12A-MBP-HBx with *Spe*I and *Nru*I in Buffer *Bam*HI (1235 bp DNA fragment).

pSFV-S2-9-pac vector The pSFV-S2-9-pac vector is non-cytotoxic, and functions as a self-replicating viral RNA whilst being maintained in the cell by puromycin selection. The pSFV-S2-9-pac vector is similar to pSFV-b12A so the cloning strategy used to generate MBP-HBx was the same as for pSFV-b12A. Further details can be found in Fig 3.1. Only the CD5-MBP-HBx DNA fragment vector could be cloned into pSFV-S2-9-pac.

3.2.3 Protein Induction and Purification

3.2.3.1 pMAL-c2 Vectors

Briefly, one litre of LB, supplemented with glucose (2 g/L) and ampicillin was inoculated with 10 mL of an overnight culture of BL21 *E. coli* cells containing the pMAL-HBx vector, and grown at 37 °C until the OD_{600} reached 0.5. IPTG was added to a final concentration of 0.3 mM to induce MBP-

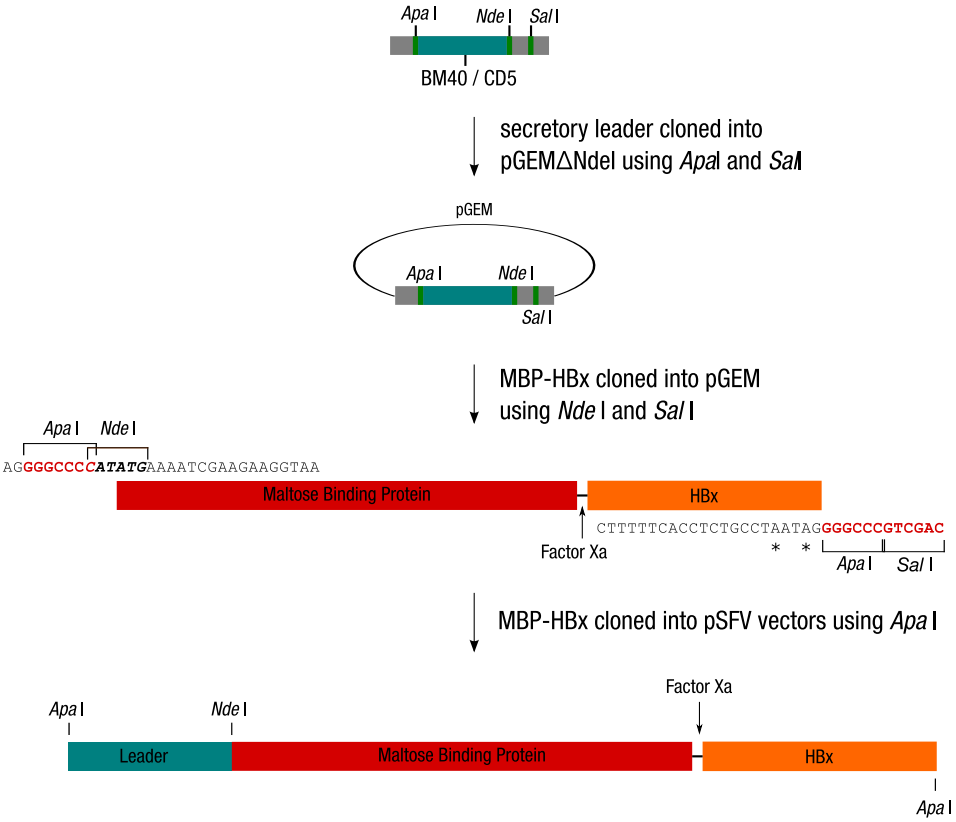

Figure 3.1: Schematic describing the cloning of MBP-HBx fusion protein, including secretory leader sequences, into the pSFV-b12a and pSFV-S2-9-pac vectors. Both CD5 and BM40 secretory leader sequences (green) were generated using complementary, partially overlapping oligos (upper panel), and cloned into a previously generated pGEMΔNdeI vector using *Apa*I and *Sal*I (upper). The middle panel shows a detailed view of the MBP-HBx fusion protein (red/orange block). The *Apa*I and *Nde*I sites are shown in red and are italicised where overlapping. The Factor Xa cleavage site is indicated with an arrow. Asterisks indicate two stop codons within HBx.

HBx protein expression. The induced culture was incubated at 37 °C for 2 hours until harvesting. Following induction, the *E. coli* cells were pelleted at 10 000 x g and 4 °C, and were resuspended at 10 mL per 1 g wet cell paste in 20 mM Tris-HCl pH 7.4, 200 mM NaCl, 1 mM EDTA and 0.1 mg/mL lysozyme and frozen at −20 °C overnight. The following day the pellets were thawed and sonicated in 25 mL aliquots for 4 pulses of 30 seconds each at 75 % power. To this lysate, 0.1% (w/v) polyethyleneimine, and 1% of streptomycin sulphate (w/v) were slowly added. The lysate was stirred for 1 hour at 4 °C to allow for the precipitation of aggregates and DNA bound proteins. The solution was then centrifuged at 37 500 x g to clarify the lysate. Protein concentration was measured by BCA assay (Pierce, IL, USA) and adjusted to 2.5 mg/mL before loading onto a 2.2 cm diameter amylose affinity resin column (New England Biolabs® Inc, MA, USA). Following loading, 12 volumes of column buffer were passed through the column to remove non-specifically bound proteins. The MBP fusion protein was eluted in 200 μL fractions, totalling 7 column volumes of column buffer supplemented with 10 mM maltose into a Pharmacia 100 fraction collector. The protein concentration of the fractions were quantified by BCA micro assay and fractions containing protein were concentrated and dialysed against a 10mM Tris-HCl pH 8.0, 25 mM NaCl buffer using a 5 kDa MWCO Amicon ultrafiltrator to a maximum of 1 mL. The sample was loaded onto a 10 mL G-75 Sepharose size exclusion column and washed with 5 column volumes of Tris buffer as mentioned above. Fractions of 200 μL were again collected and tested using the BCA micro assay. Fractions containing protein were pooled and concentrated with a 10 kDa MWCO Amicon ultrafiltrator to at least 1 mg/mL. The concentrated protein was stabilised with 50% (v/v) sterile glycerol and snap frozen in liquid N_2 for storage at −70 °C or used in a cleavage reaction with Factor Xa.

3.2.3.2 Cleavage of MBP-HBx using Factor Xa

The cleavage of HBx from the fusion protein was achieved by incubating purified MBP-HBx with Factor Xa (New England Biolabs, MA, USA) at a 50:1 (w/w) ratio. The optimal cleavage time required was assessed by splitting the reaction into equal volumes and incubating at room temperature. After 24, 48, and 72 hours, SDS-PAGE sample buffer was added to a reaction to a 1x concentration and the sample was boiled for 5 minutes at 95 °C. The samples were assessed using SDS-PAGE and Western blot.

3.2.4 *In Vitro* Transcription (IVT) and Electroporation of SFV Vectors

Each recombinant expression vector, including the split-helper plasmid used to generated SFV virions, was *in vitro* transcribed as follows. Five μg of DNA was linearised with *Spe*I in a standard reaction at 37 °C and confirmed on a 0.8 % agarose gel. Linearised DNA was purified using phenol:chloroform (see Appendix B.2.2) and resuspended in RNase free dH_2O at 1 μg/μl. A 50 μl IVT

Table 3.1: Components of a standard IVT reaction for the generation of capped SFV mRNA.

Reagent	Volume (μl)	Final Concentration
SP6 Buffer	5	1x
rNTPs	25	4mM A/C/UTP 0.8mM GTP, 6.4mM G(5′)ppp(5′)G
Pyrophosphatase	2	0.004 units/μl
$MgCl_2$ (500mM)	1.4	14 mM
linear DNA	2	2 μg
dH_2O	12.6	
Final Volume	50	

reaction was set up for each of the plasmids as shown in Table 3.1 and incubated at 40 °C for 2 hours.

BHK-21 mammalian cells were were seeded to 80% confluency in a T75 flask less than 24 hours before electroporation. The medium was removed and cells were washed with sterile PBS without Ca^{2+} or Mg^{2+}. Following this, 2 mL of 1x trypsin-PBS-EDTA solution was added, and cells were incubated for 1-2 minutes until they started to detach. Ten mL of BHK-21 medium was added and cells were resuspended to obtain a single cell suspension. The cells were transferred to a 15 mL centrifuge tube, and pelleted by centrifugation at 129 x g for 3 minutes. The cells were washed with 15 mL sterile PBS, and re-centrifuged. This process was repeated twice. Cells were re-suspended in PBS, counted using a haemocytometer and adjusted to 10^7 cells/mL with PBS. Fifty microlitres (50 μl) of IVT mRNA was added either singly or premixed together with the 2 split helper mRNAs into a 0.4 cm electroporation cuvette. The cuvette was pulsed twice with a 15 second interval, at 850 V and 25 μF at room temperature. Cells were immediately transferred into pre-warmed BHK-21 media and incubated in a humidified atmosphere with 5% CO_2, at 33 °C for 40-48 hours to maximise generation of virus.

3.2.5 Harvesting of Viral Particles

The medium was aspirated and centrifuged at 40 000 x g for 30 minutes to remove cells and cellular debris. The supernatant was filtered through a 0.45 μm filter before being transferred to an ultracentrifuge tube 25 x 89 mm (Beckman Coulter, CA USA,). A 5 mL cushion of 20% sucrose was gently pipetted to the bottom of the tube using a glass Pasteur pipette so as not to mix the two layers. Tubes were then centrifuged for 90 minutes at 82 705 x g and 4 °C in a vacuum to pellet the

virus. The supernatant was aspirated without touching the bottom of the tube. Five hundred μl of TNE buffer was added to each tube, covered with parafilm and incubated at 37 °C to resuspend the viral pellet. The concentrated stock was sterilised by filtration with a 0.22 μm filter, aliquoted and stored at -70 °C.

3.2.6 Viral Titering by Immunofluorescence Microscopy Imaging

To titre SFV viral stocks, BHK-21 cells were grown to 80 % confluency in complete BHK-21 medium. Cells were seeded into 6 well plates with sterile glass coverslips, at a cell density of 5 x 10^5 cells/ml. Cells were left to grow at 37 °C for 8 to 16 hours. A dilution series of viral stocks was made between 10^{-2} to 10^{-5} in MEM (Gibco with 0.2 % (w/v) BSA, 2 mM L-glutamine, 20 mM HEPES). The cells were washed twice with PBS, thereafter 500 μl of virus in minimal media was added to the cells. The cells were incubated at 37 °C for 1 hour, and the flask was rocked every 10 minutes. The virus-containing medium was removed and replaced with fresh medium, after which the cells were incubated overnight at 37 °C.

Cells were fixed and labelled with antibody as described in section 2.2.7.1, with the exception that the primary antibody was rabbit antiserum specific for SFV replicase, diluted to 1:400. The secondary antibody was a Cy5 or FITC-conjugated goat anti-rabbit antibody.

3.2.7 HBx Activity Assay

To assay purified recombinant HBx for functional activity, the transactivation of the interleukin 8 (IL8) promoter by HBx was tested[176] using a HeLa cell line stably expressing eGFP from an IL8 promoter (kindly provided by Dr Musa Mhlanga CSIR, South Africa). HBx function was measured as a change in eGFP expression within these cells.

Stable HeLa cells were seeded at 80% confluency onto glass coverslips before the addition of : 1) no protein (background control), 2) TNFα (positive control), 3) MBP (negative control), 4) MBP-HBx (uncleaved), 5) MBP-HBx (cleaved). Purified proteins were added to a final concentration of 100 μ/g mL. Cells were fixed at 2 and 24 hours according to section 2.2.7.1 and were mounted with VectaShield containing DAPI (Vector Laboratories, CA, USA) to stain the nuclei.

Following mounting, 10 images of each sample were acquired in both the DAPI and GFP channels by blind sampling i.e. image acquisition was acquired in the DAPI channel to prevent sampling bias in the GFP channel. Mean fluorescence intensities of the GFP channel were analysed using ImageJ image analysis software[177]. Significant fluorescence intensity differences between samples was determined by applying a Students-t test in the R statistical suite. The immunofluorescence microscopy facilities at the Gene Expression and Biophysics Group, CSIR Biosciences, Pretoria South Africa, were used to acquire the images.

3.3 Results

3.3.1 Expression of MBP-HBx fusion protein in *E. coli*

As shown in the elution profile of MBP-HBx from the maltose resin, (Fig 3.2), and the elution profile from the G-75 size exclusion resin (Fig 3.3), a clean preparation of MBP-HBx was achieved. The preparations of the fusion protein showed a certain degree of instability as indicated by the degradation products on a Western blot (Fig 3.4-A).

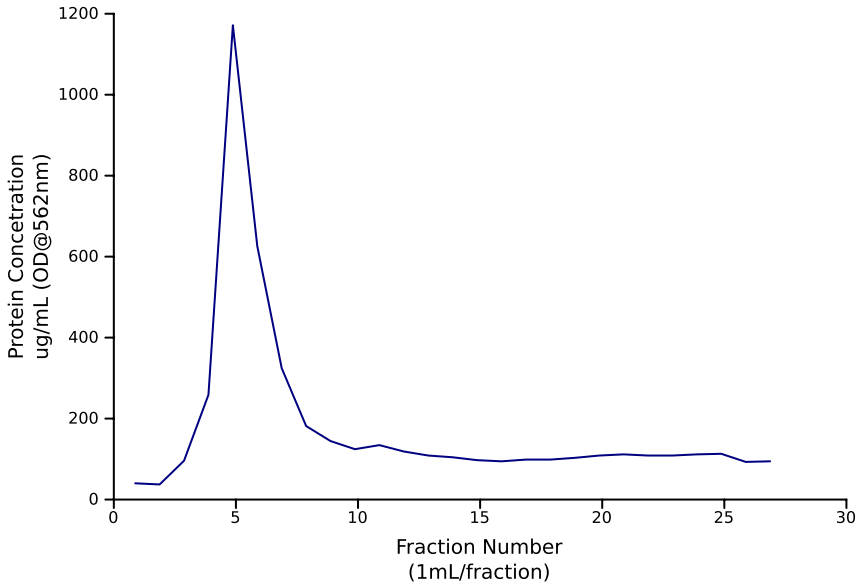

Figure 3.2: Affinity elution profile of *E. coli* expressed MBP-HBx eluted with 10mM maltose, after washing for 20 column volumes. Fractions were 200 μl each and those containing protein as measured by BCA assay were pooled for subsequent storage or cleavage. A standard curve was generated against a 0-2000 μg/mL range of BSA. A single peak correlating with MBP-HBx protein eluted in fractions 3 to 7 suggested a clean preparation was achieved.

3.3.2 Cleavage of MBP-HBx fusion with Factor Xa

Cleavage of HBx from the MBP tag was found to be more reproducible if performed directly after purified and concentration protein. As reported by Liu and colleagues[128], cleavage of HBx off MBP-HBx did not increase appreciably after 48 hours as shown in Fig 3.4-B. The cleaved products were separated using G-75 size exclusion chromatography and fractions found to contain HBx were concentrated and snap frozen with 50 % glycerol in liquid N_2 and stored at $-70\,^{\circ}$C.

3.3.3 *In vitro* activity assay of HBx protein

Previous studies and work from our laboratory have shown that recombinant *E. coli* expressed HBx, after denaturation and refolding, possesses biological activity[95,105]. As HBx possesses no

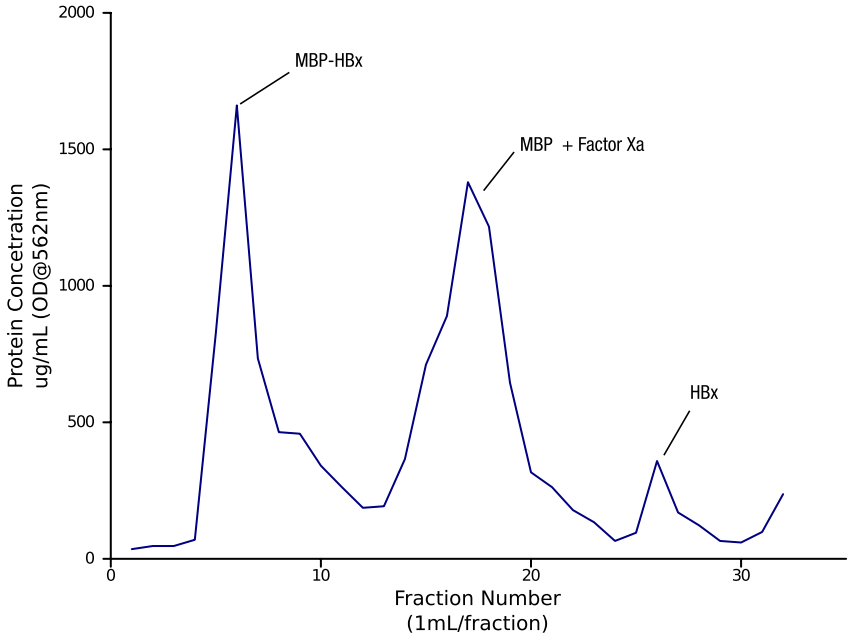

Figure 3.3: G-75 Sephadex size exclusion elution profile of *E. coli* expressed MBP-HBx after 48 hours of cleavage with Factor Xa. Fractions were 200 μl each and those containing protein as measured by BCA assay were pooled for cold storage. A standard curve was generated against a 0-2000 μg/mL range of BSA. Three peaks correlating with HBx alone, MBP + Factor Xa and MBP-HBx were observed. The separation of the three peaks permitted easy isolation of purified HBx.

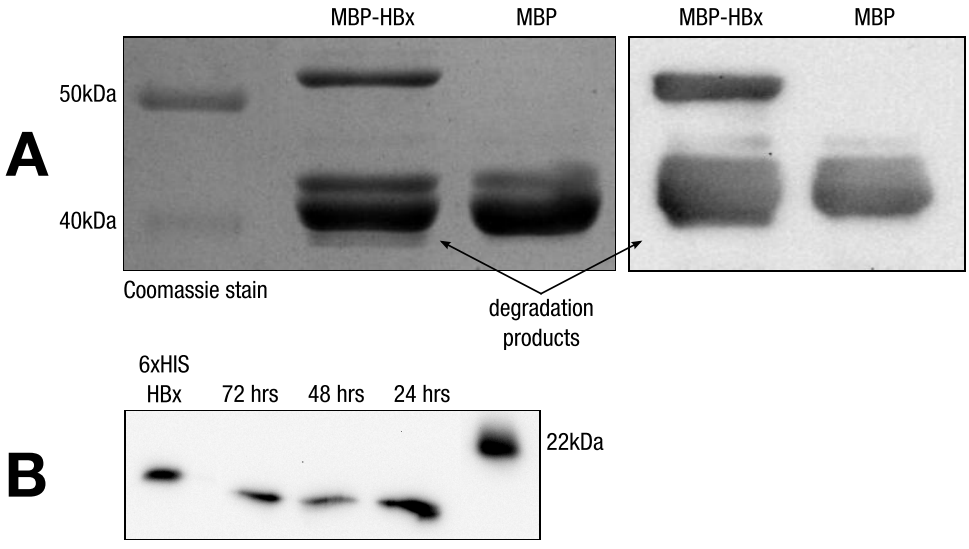

Figure 3.4: Western blot and Coomassie stain of (A) MBP-HBx and (B) cleavage of HBx from MBP (B). **A** - MBP-HBx and MBP preparations after affinity and size exclusion chromatography showed a high degree of purity. Degradation products are indicated with arrows. **B** - Western blot using the anti-HBx 6.1 antibody to detect HBx from a time cleavage assay of MBP-HBx using Factor Xa. Samples were taken at 24, 48 and 72 hours. No increase in cleavage was seen after 48 hours.

enzymatic activity, its ability to transactivate promoters serves as a measure of biological activity. Typically assays that measure changes in Chloramphenicol acetyltransferase (CAT) levels are used[64,178–180]. CAT is expressed from the HIV LTR promoter and a doubling of protein in the presence of HBx can be expected. In this work, HeLa cells stably expressing eGFP off an IL8 promoter were used to assay the transactivation activity of HBx. As shown in Fig 3.5, HBx can transactivate IL8 leading to an increase in GFP expression. Interestingly, HBx seemed to retain its activity whilst attached to MBP. Importantly the negative controls including MBP alone, and no added protein showed no increase in GFP levels.

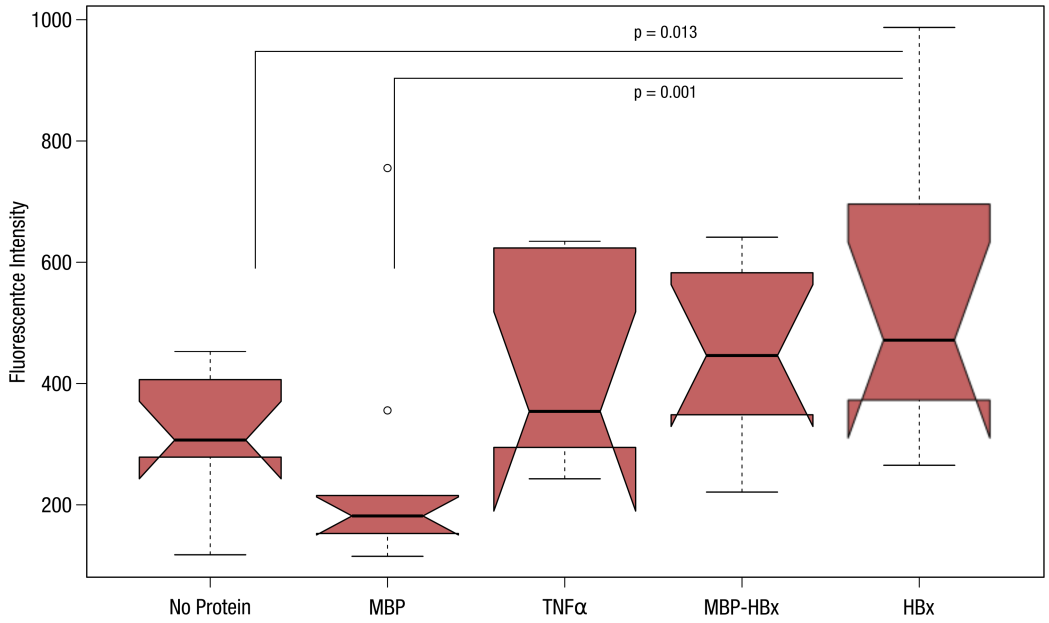

Figure 3.5: Notched box and whisker plots of comparative fluorescence intensities (n=10) of IL8-GFP containing HeLa cells after exposure to either MBP, TNFα, MBP-HBx or HBx. The difference between the fluorescence values of HBx versus MBP and HBx versus No Protein samples were determined by applying a Students-t test with a 95% confidence level. The results show that MBP-HBx and HBx alone caused an increase in eGFP fluorescence intensity similar to the TNFα positive control. The notches on the box represent the confidence intervals for the medians. If notches of two medians do not overlap the medians are significantly different at a 95 % confidence interval. A key explaining a box and whisker plot is presented in Appendix A.7.

3.3.4 Expression of MBP-HBx using pSFV-b12a

The expression and partial secretion of the MBP-HBx fusion protein using the pSFV-b12a vector was successful, showing that this is a promising construct for mammalian production of recombinant HBx. The Western blot in Fig 3.6 indicates that the majority of the protein was located intracellularly (Lane 3), although comparable amounts of protein was secreted into the growth medium (Lane 1). Further attempts at expanding the use of this vector through the generation of virions, were

unsuccessful (data not shown). Conditions attempted included lowering the incubation temperature to 33 °C, increasing the amount of mRNA electroporated into cells from the pSFV-b12a and two helper plasmids, and optimising the timing of electroporation to less than 18 hours after splitting the cells. Despite these attempts, titres of virus were never more than $10^{4\text{-}5}$ infectious units (IU) per mL. As several log increases to $10^{8\text{-}9}$ IU were needed for large-scale infections, the use of this vector was not pursued.

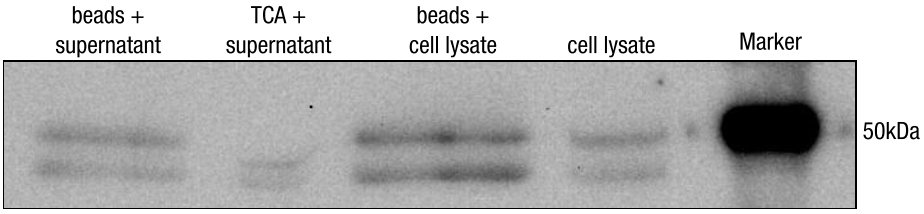

Figure 3.6: Western blot of protein samples taken from cells electroporated with pSFV-b12a-ppt-MBP-HBx mRNA. MBP-HBx and degradation products (lower bands) were detected in all samples analysed. Lane 1 shows beads used to purify MBP from supernatant (control for presence of MBP), Lane 2 shows TCA plus supernatant (non-specific protein precipitation). Lane 3 shows beads used to isolate MBP from cell lysate. Lane 4 shows cell lysate only.

3.3.5 Expression of MBP-HBx using pSFV-S2-9-pac2A-CD5-MBP-HBx

During infection with the pSFV-b12A vector, cellular processes are sequestered by the demands of the SFV RNA during recombinant protein production[181]. The cytotoxicity of this vector is such that after 72 hours all infected cells are dead and the electroporation or infection needs to be repeated in fresh cells. To allow for continuous protein expression and purification as opposed to the repeated transient expression of pSFV-b12A vector, the use of the non-cytotoxic pSFV-S2-9-pac2A vector was pursued. This vector circumvents the need to produce virions for infection, and instead relies on the replication of the SFV mRNA within the cytoplasm of the electroporated cell. To maintain mRNA presence, the replication of the viral mRNA is directly linked to a puromycin selection marker. This vector allows positively selected cells to be grown to a large-scale. Three attempts at establishing pseudo-stable BHK-21 cells expressing MBP-HBx were unsuccessful. Western blots aimed at detecting MBP or MBP-HBx in the supernatant of electroporated cells were negative (data not shown). This was true for the SFV nsp2 protein, which is part of the SFV replicase and should be detected in cells containing an SFV vector (Fig 3.7 Lane 1 and 2).

Nsp2 was however detected in the positive control derived from the pSFV-GFP vector (Fig 3.7 Lane 2). The BHK-21 cells with no vector showed an unusual tolerance for puromycin at the 5 μg/mL concentration reported to kill cells not expressing a puromycin resistance cassette. Although the concentration of puromycin was increased from 5 to 7 μg/mL, this did not result in the selection

of MBP-HBx expressing cells. Clumps or colonies of BHK-21 cells are known to be resistant to puromycin and so care was taken to ensure that cells were split at a low confluency and were not permitted to overgrow. It is possible that the BHK-21 cells used in these experiments had greater than normal interferon activity which would cause the SFV vector to be rapidly expelled from the cell. However, this was not determined.

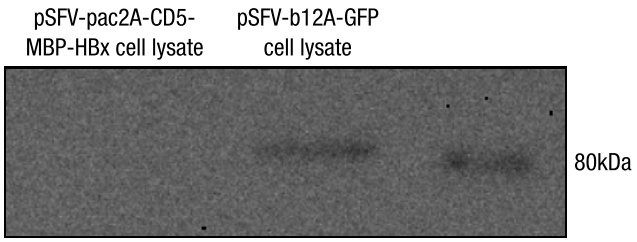

Figure 3.7: Western blot of pSFV-S2-9-pac2A-CD5-MBP-HBx to screen for the presence of non-structural protein 2 (nsp2) protein from pSFV-S2-9-pac2A-CD5-MBP-HBx containing cells. The nsp2 protein indicates the presence of an SFV vector (positive control) and serves as confirmation of the presence of an SFV RNA replicon. Nsp2 protein was only detected in the positive control (pSFV-b12A-GFP) sample.

3.4 Discussion

The expression and purification of soluble and native HBx using a MBP fusion approach was a major achievement in expressing a protein renowned for its insolubility and temperamental nature. As researchers can now produce preparative scale quantities of this protein, at least in *E. coli*, attempts at elucidating its structure, which for so long have been frustrated, are now more likely to be met with progress.

The MBP system used in this work, whilst successfully producing soluble, functional HBx, could be further improved to increase the final yield of HBx. A major concern is the considerable quantity of degradation products from the MBP-HBx fusion protein that complicates the purification procedure (Fig 3.4-A). The complication lies in that degradation products compete for binding on the maltose resin, thus necessitating an additional separation step using G-75 size exclusion or Anion exchange chromatography. Furthermore, and more importantly, the degradation products suggest that the yield of HBx from the fusion protein is continually decreasing. The standard purification methodology suggested by New England Biolabs and used in this work are column dimensions of 2.5 cm, which dictates a maximal flow rate of 1 mL/minute [10 x (diameter of column in cm)^2]. This flow rate holds true for both the loading of the column and the washing and elution of any bound proteins. A further limitation is that the maximum loading protein concentration must be approximately 2.5 mg/mL to avoid losses, channelling and blocking of the column resin. To reduce the load of protein applied to the column, affinity precipitation of MBP-HBx was attempted[182], although this was not included here. Cationic starch (part of affinity precipitation) is not a standard, readily available, nor often used laboratory chemical, and so differences in the cationic starch qualities to those described were likely to be responsible for the negative results. Improvements to the protocol and reagent availability of this initial affinity precipitation technique could greatly reduce the time and cost to purify MBP tagged proteins.

Without the initial affinity precipitation, the MBP-HBx fusion protein is on the column for 14 hours or more. Loading of the maltose column takes 4-5 hours depending on lysate volume and protein concentration, and washing of the column occurs over an additional 10-14 hours. If the MBP tag was altered to include a rapidly purifying 6xHIS tag as an initial step, the purification of the protein could be greatly accelerated. The MBP affinity resin could then be used as a final polishing step as opposed to its current use as an initial purification step.

Further increases in the stability of the MBP-HBx fusion could be achieved if an additional MBP tag was attached, possibly in tandem, to the fusion protein. This had a positive impact on another difficult to express protein, BRCA2[174], where two N-terminal MBP tags were attached. The downside of this is that increased stability of the fusion protein is counteracted by an increase in the proportion of MBP in the fusion, and a subsequent decrease in the final yield of the protein of interest (e.g. a 2

N-terminal MBP-MBP-HBx protein would decrease the final yield of HBx from 28% to 17%). It would be important to determine whether an increase in stability as a result of additional MBP tags results in a real increase in the protein of interest.

Alternatively, other solubility enhancing tags like NusA (*E. coli* transcription termination anti-termination factor) or SUMO (Small Ubiquitin-like Modifier) could also be tested, either alone or in combination with each other[183]. Both of these tags have been shown to have similar solubility promoting effects on their fusion partners. Further attempts should focus on keeping the fusion cleavage position site as close to the beginning or end of HBx as possible to ensure that there are no modifications to the target protein sequence or structure. Additionally, other secretory leaders (prepro-trypsin or BM40) that resisted subcloning from the pGEMΔNdeI plasmid (data not shown) could be tested for improved secretory performance.

The MBP-HBx fusion protein was sub-cloned into the SFV protein expression system as these vectors can be used for large scale expression of proteins[184]. Notably, reports of infections of 1-10 litres of suspension cells, producing around 1-20 mg/L of protein, make this system useful for proteins amenable to this mode of production. The system does have several shortcomings which have been addressed comprehensively over the years. Transient infections of cells by either the viral mRNA or viral particles results in a complete hijacking of the cellular machinery, so much so that infected cells are dead by 72 hours. This has been ameliorated by the creation of SFV vectors with lowered cytotoxicity[185], and more recently a non-cytotoxic, self replicating SFV vector, used in the experiments here (pSFV-S2-9-pac vector)[170,186]. Additional concerns, although admittedly beneficial for experimental purposes, are the ability of alphavirus vectors to infect a wide variety of cells and hosts (e.g. avian, insect and mammalian cells). This represents a significant biosafety problem which numerous researchers have addressed. Early use of SFV showed the production of replication-proficient viruses (RPVs) which presented a danger of infecting the experimenter or other lab personnel. This has largely been overcome by separating the viral helper and vector RNA, preventing a productive infection. Recent enhanced biosafety modifications include the split-helper system whereby the capsid and spike helper regions are split into two independent RNAs[173]. An additional mutation abolishing capsid self-cleaving activity increases the biosafety of this system even further[187,188].

Despite the expression improvements and biosafety modifications made to the SFV system, the generation of substantial quantities of MBP-HBx was still not successful. Difficulties in *in vitro* transcription and electroporation of single or multiple 12 kb mRNAs were largely overcome, but generating high titres of viral particles and/or increasing packaging efficiency was not improved. The selection of pseudo-stable cells using the pSFV-S2-9-pac2A-CD5-MBP-HBx was also not successful despite multiple attempts at selection. The BHK-21 cells used showed a remarkable resistance to puromycin, that did not prevent SFV mRNA negative cells from expanding.

The absence of SFV mRNA in the cells was deduced from the absence nsp2 proteins in Western blots, as compared to the controls (Fig 3.7 Lane 1 and 2). It was not determined exactly why the generation of high viral titres and the selection of stable cells was not successful, but it is possible that the BHK-21 cells used produced elevated levels of interferon which induced an antiviral state in the cells thus expelling the SFV RNA[189].

Despite the shortcomings and technical difficulties of working with SFV based systems, SFV was successfully used to express MBP-HBx protein as well as permitting its secretion from infected BHK-21 cells. The MBP-SFV expression system presented in this chapter, represents a significant milestone in HBx research and is an excellent point from which to pursue preparative scale production of soluble, eukaryotic HBx. The system demonstrated that with additional modifications to the SFV expression construct and closer collaboration with researchers actively developing these SFV vectors, significant progress could be made in the large scale expression and structural determination of HBx. This will ensure that structural data can be gleaned and will assist researchers in elucidating the genuine biological role of HBx in hepatocarcinogenesis.

Due to a lack of structural studies related to the unavailability of eukaryote expressed HBx, the genuine biological role of HBx in hepatocellular carcinogenesis, as well as its roles in the lifecycle in HBV remain controversial. Thus eukaryotic expression of HBx using the systems presented here would be highly beneficial. In addition, studies on the overlapping genes of the HBV genome would provide interesting insights into HBx. Evolutionary analysis of HBV sequences, would provide interesting insights into the evolutionary pressures experienced by overlapping and non-overlapping regions. Importantly, this analysis could reveal information on regions of biological importance which could assist future structural studies. Additionally the elucidation of sites that are prone to mutation could assist future antiviral therapies by directing researchers away from unsuitable regions. Furthermore, as HBV subtypes are still largely separated geographically, analysis of selection on the subtypes could reveal information on the evolution of HBV in different geographic regions. The evolutionary analysis of HBx is dealt with in Chapter 4.

Chapter 4

Evolutionary Analysis of Hepatitis B Virus HBx Protein

4.1 Introduction

Overlapping genes are created as a result of substitutions occurring within an upstream codon in an adjacent gene. This substitution results either in the addition of an initiation codon or the removal of a termination codon. The reading frame of the downstream gene is thus extended proportionally. This mechanism is known as over-printing[190]. Overlapping sequences are constrained in their ability to adapt and this effect may imprint on the remainder of the gene. Despite this constraint, it is proposed that overlapping genes offer a fitness advantage to viruses as the assemblage of proteins is increased without increasing the genome size[191,192] as this in turn is constrained by packaging capacity of the capsid[193]. Additionally, several analytical models have shown that overlapping sequences may palliate the effects of detrimental substitutions, especially when mutation rates are high[194,195]. This may be the case with HBV, although a weak host immune response results in the viral genome being relatively stable[196]. Overlapping genes reduce redundancy (called antiredundancy) in genomes and this seems to be a characteristic of large populations such as viruses. Antiredundancy serves to remove mutants, thereby improving the robustness of wildtype genomes[197].

When an overlap is created, the newly added sequence that was previously non-coding has not undergone functional adaptation and thus mutations will typically have no fitness cost. This ensures these mutations are maintained within the population. In general, when an overlap is created, the additional protein sequence is disordered and is taxonomically restricted[198].

Hepatitis B Virus is characterised by uni-directional overlapping of all of its genes such that the $3'$ end of one gene, overlaps with the $5'$ end of another gene (Fig 4.1). The *hbx* and *core* genes occupy a 123:312 phase with respect to the *polymerase* gene such that the first codon base in the

hBx or *core* gene corresponds to the degenerate third position of *polymerase*, whereas the *surface* protein occupies a 123:231 phase with respect to *polymerase*. The degree of overlap within the HBV genome is considerable. The genome size is 3.2 kb and can encode 180 kDa of protein, as opposed to a theoretical maximum of 120 kDa for a non-overlapping 3.2 kb genome[199]. HBV regulation relies on the interactions of several *trans*-acting factors, and their *cis* motifs in overlapping regions adds further evidence to the efficient organisation of the HBV genome (reviewed in[11]).

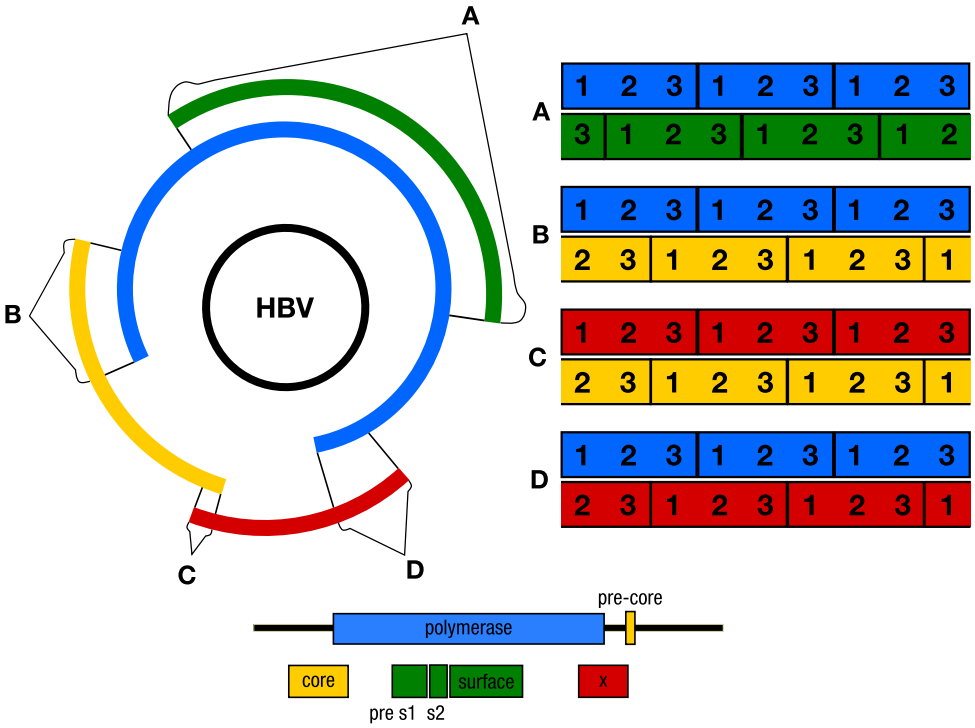

Figure 4.1: Representation of the overlapping genome of HBV and the reading frames associated with each overlap. Both linear and circular representations of *core* (yellow), *polymerase* (blue),*surface* (green) and *hbx* (red) are shown with each of the reading frames labelled as 1,2 or 3. (A-D) the first nucleotide position of *polymerase* corresponds to the third nucleotide position of *surface* (p1/s3), the second nucleotide position of *core* (p1/c2), the second nucleotide position of *hbx* (p1/x2), whereas the third nucleotide position of *hbx* corresponds to the first position of *core* (x1/c2).

In Chapter 2 (see Fig 2.14) it was suggested that HBx possesses disorder, additionally it is known to be restricted to the genera *orthohepadnaviruses* which only infect mammals. This indicates that *hbx* evolved *de novo* via the over-printing mechanism mentioned above. This is also true of unspliced RNA viruses (e.g. tymovirus and luteovirus) where overlapping genes have been shown to code for proteins that are either partly or wholly disordered, have an unusual sequence composition and play a role in viral pathogenicity or spread[192,198].

At a sequence level, *hbx* can be divided into 3 distinct segments (Fig 4.1). The first segment overlaps with *polymerase* on its N-terminus (54 % of *hbx* length), a non-overlapping segment in the

middle of *hbx* (40 % of *hbx*), and a short C-terminus segment that overlaps with the *pre-core* gene (3.9 % of *hbx*). The evolutionary pressures on these three sections of *hbx* are hypothesized to be different from each other, since the position of a nucleotide substitution strongly influences whether that substitution will be synonymous in the other reading frames or not. A mutation in: 1) the first codon position causes a non-synonymous mutation in 60 out of 64 cases (93 %), 2) the second position affects 63 out of 64 cases (98 %) and 3) the third position affects only 16 out of 64 cases (25 %, amino acid redundancy). Therefore a substitution in the first position of *hbx* will probably cause an amino acid change in *pre-core* (position 1 of *hbx* is position 2 of *pre-core*, x1/c2 - compare *hbx* in red to *pre-core* in yellow in Fig 4.1). Furthermore, a substitution in the first position of *hbx* will probably result in an amino acid change in HBx, but not in Polymerase, x1/p3 (compare *hbx* in red to *polymerase* in blue in Fig 4.1).

Despite the fact that 1) *hbx* overlaps with *pre-core* and *polymerase* in the HBV genome; 2) *hbx* contains a disordered region and 3) *hbx* may be involved in the pathogenicity of the virus, no studies have been published to determine the types of selection within *hbx*. Thus selection within the HBV genome and within *hbx* in particular was assessed here. It was hypothesized that the evolution of *hbx* may mirror the research done by Zaaijer *et al*[200] where the evolution of the HBV *surface* and *polymerase* genes occurred at s2/p3 and p1/s3 respectively, meaning that substitutions at s2 and p1 (non-synonymous) probably did not result in an amino acid change at p3 and s3 respectively.

Additionally, we hypothesized that the negative regulatory domain, with its predicted disordered segment (see Fig 2.14), would be under different selection compared to the transactivation domain. Considering that HBx has been reported to play a role in the progression of chronic HBV infection to liver cancer, this work also aimed to provide direction with regards to the development of chemotherapeutics or inhibitory small molecules through the application of phylogenomics.

4.2 Methods and Materials

4.2.1 Sequence Extraction and Analysis

To determine the evolutionary patterns of the 4 main genes of HBV (*hbx, surface, core, polymerase*), the variation in the codons and amino acids of the 8 genotypes were gathered as follows: the complete list of 2475 separate HBV genomic sequences was downloaded from Genbank and re-genotyped using the Oxford HBV Subtyping Tool[201,202]. This was to ensure that sequences had been correctly annotated. The type sequences within each genotype were duplicated by concatenation to facilitate correct sequence alignment, and then used to extract the 4 genes from the database of sequences using the Waterman alignment in the EMBOSS suite of tools[203]. Gene sequences containing insertions or deletions greater than 10 nucleotides were removed from the datasets as this affected the computations. All identical sequences were removed as this affects the convergence of the evolutionary models. All stop codons were removed, as well as sequences containing premature stop codons as these are not tolerated by the analysis software. This was in line with the requirements for downstream analysis.

Overall, the dataset of sequences covering all eight genotypes was too large to be computed in a reasonable time. Thus, a random sampling of a maximum of 75 *polymerase* sequences per genotype was analysed (due to the sequence length). Approximately 200 sequences for each of the *hbx, surface* and *core* genes were sampled. All sequences were extracted using Python code as described in Appendix C.

The extracted gene sequences were translated and aligned using MUSCLE[204]. Initial phylogenetic trees were calculated using the BioNJ function[205] of SeaView software[206] with a Jukes-Cantor correction and 500 bootstraps. Protein sequence alignment was included simply to ensure an inframe alignment. All subsequent analyses were performed on the gene sequences only, unless specified otherwise.

The CODEML module of PAML 4.4d[207,208] was applied to the codons to estimate the relative rates of synonymous and non-synonymous nucleotide substitutions. The *hbx, core polymerase* and *surface* genes were analysed by maximum likelihood (ML) approximation separately but concurrently. The nested site models (0, 1 and 2, 7 and 8) of PAML were run on the cluster supported by Wits Bioinformatics running Scientific Linux 5.4. Computations of datasets amounted to several thousand hours in total for all four genes.

In cases where the nested site models were not converging as expected, an initial input phylogenetic tree from model 1 was used to overcome parameter trapping and increase the speed of analysis. Likelihood-ratio tests (LRT) and Bayes Empirical Bayes (BEB) statistics[209–211] were applied as described in the PAML manual (http://abacus.gene.ucl.ac.uk/software/paml.html).

4.2.2 Calculation of Codon Position Substitutions using Entropy

The calculation of entropy reflects the amount of variability through a column in an alignment and is a measure of the lack of "information content" at each position in the alignment. In other words, it is a measure of the lack of predictability at a particular position in the alignment. Entropy calculations for each of the three codon positions of *hbx* were performed using the Entropy function of BioEdit software version 7.0.5.3. (http://www.mbio.ncsu.edu/bioedit/bioedit.html)[212]. Calculations for the three reading frames tested (x1/p3, x2/p1, x3/p2) were performed by removing one nucleotide and recalculating the entropy for the reading frame.

4.2.3 Branched Site Models

To determine the selective pressures occurring in *hbx* in the different genotypes of HBV, all *hbx* sequences were combined and aligned using MUSCLE. The CODEML module of PAML 4.4d was again used and Model 2, NSsites 2, Method 1 and Model 2, NSsites 3, Method 1 were applied to the data set. An initial phylogenetic tree was generated using BioNJ with a bootstrap of 100[205]. All branches within each genotype were labelled as a clade using "$1". The branched site models were applied to each genotype consecutively.

4.2.4 K_a/K_s estimations

The ratio of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site is represented as K_a/K_s. The ratios were calculated using SWAAP 1.03[213] with the Nei-Gojobori distance estimation method (sliding window length = 15 nt; window step = 3 nt)[214]. Sequences of each *hbx* genotype were subdivided into segments of 99 sequences each and analysed separately (due to a limitation on number of sequences that could be entered into SWAAP). Results were combined, giving a total of 1666 comparisons.

4.3 Results

4.3.1 Entropy Calculations for *hbx* Reading Frames

The *hbx* gene can be divided into three sections based on the overlap with *polymerase* and *pre-core* (Fig 4.1). The entropy for each of these sections (overlap with *polymerase*, overlap with *pre-core* and non-overlapping region) was calculated, in three different reading frames. The results shown in Fig 4.2 indicate that where *hbx* overlaps with *polymerase* (amino acids 1-83), the variability in *hbx* is primarily non-synonymous and occur at x1/p3, while the same mutations are synonymous in *polymerase*. The non-overlapping section of *hbx* shows a high proportion of evolution occurring by non-synonymous substitutions in the x1 reading frame, and similarly in the x2 reading frame. This suggests that the lack of overlap with additional sequences permits evolution in this *hbx* section to be mainly non-synonymous. Regulatory elements like the basic core promoter which modulates the precise initiation of both the pre-core and pregenomic RNA are present in this section. There are also sequence motifs located upstream and within the basic core promoter which have been shown to interact with several liver-specific transcription factors. Thus the evolution in this section of *hbx* is not only directed to the *hbx* sequence, but the regulatory sequences contained therein. Whether this represents a constraint or not was not determined.

The section of *hbx* that overlaps with *pre-core* translates to only eight amino acids and thus there is less data for entropy analysis. Nonetheless, the data suggested that where substitutions occur in the x2/c3 and x3/c1 reading frames, the alternate gene is more likely to undergo a synonymous mutation.

4.3.2 K_a/K_s Estimates of Positive Selection

To determine whether the negative enhancer and trans-activation domains of HBx are under positive selection, the use of a sliding window over the whole of HBx was used to calculate K_a/K_s ratios for each amino acid site. A value less than one indicates negative selection, a value equal to one indicates neutral selection and a value greater than one indicates positive selection at a particular site.

In Fig 4.3 the K_a/K_s ratio for each sliding window is compared between all genotypes (e.g. sliding window one which corresponds to the first 15 amino acids of genotype A is compared to the same amino acids in all the other genotypes, and similarly for the remaining sliding windows).

The majority of data points for genotypes A-E in Fig 4.3 cluster between zero and one which is indicative of negative selection. However, there are several values substantially higher than one, indicating positive selection at certain sites within HBx. Only genotype E showed a lower clustering of K_a/K_s ratios , thus indicating fewer sites under positive selection and higher negative selection.

The K_a/K_s ratios of genotypes A-F were also plotted over the length of the gene to give an

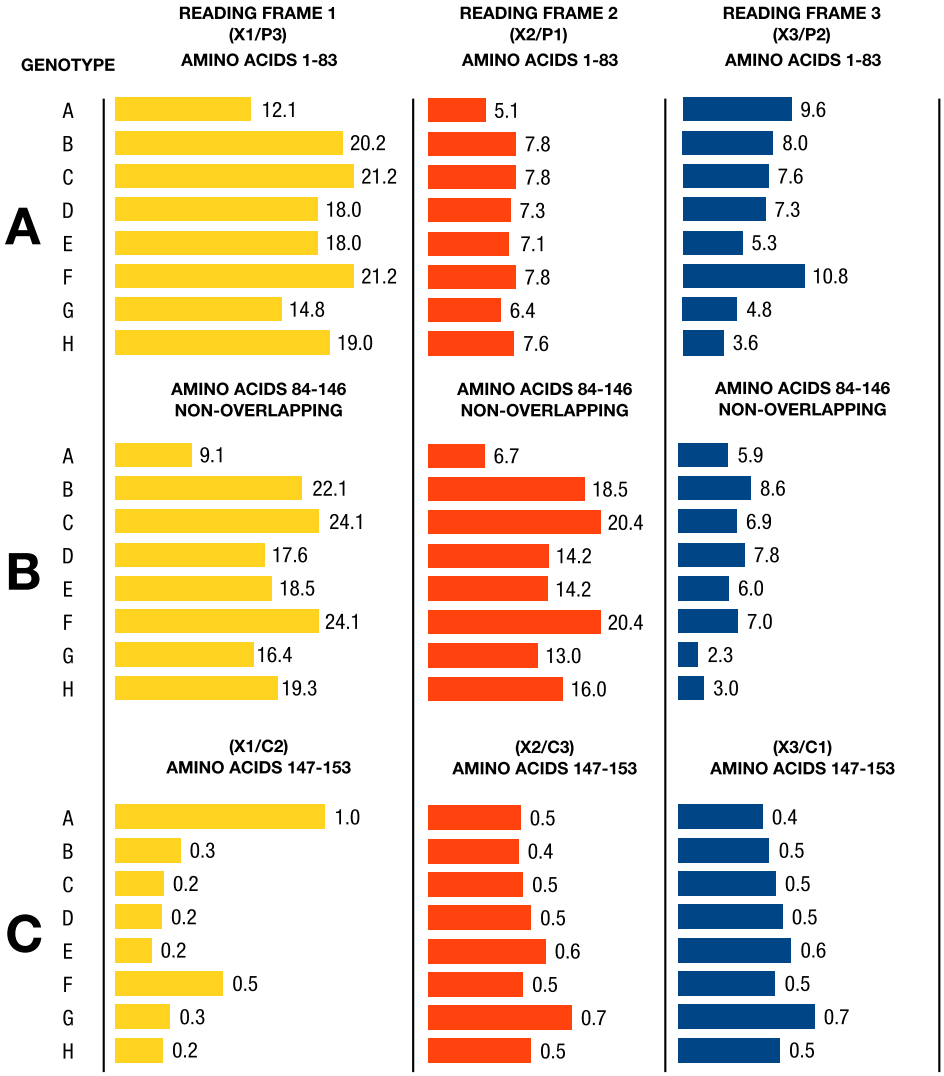

Figure 4.2: Split bar matrix chart representation of the cumulative entropy determined for HBx, in all genotypes, and for all three codon positions of *hbx*. The graph is divided up into 3 sections: Panel A, corresponds to the overlap of *hbx* with *polymerase* (amino acids 1-83), Panel B corresponds to that portion of *hbx* not overlapping with HBV genes (amino acids 84-146), and Panel C, where *hbx* overlaps with *pre-core* of the Core protein (amino acids 147-153). In all panels, the x1 position is shown in yellow, the x2 position in red, and the x3 position in blue. Entropy values are provided at the end of each line. Results in panel A, show that the majority of substitutions occur at x1 (non-synonmous), which coincides with p3 (synonymous). Panel B shows an increase in x2 substitutions as there is no constraint from any overlapping gene. Panel C shows almost identical levels of substitutions at x2/c3 (non-synonymous) and x3/c1 (synonymous). Comparisons are only made across a panel and not between panels.

indication of which regions are under positive selection across the different genotypes (Fig 4.4). Overall there are clusters of positive selection (indicated by K_a/K_s ratios higher than 1), with two spikes in the negative regulation domain (nucleotides 0-200) at positions 0-40 and 80-160. The transactivation domain (nucleotides 200-450) shows 3 spikes around nucleotides 300, 375 and 440. Overall these results suggest that the negative regulation domain of *hbx* in under positive selection across all genotypes.

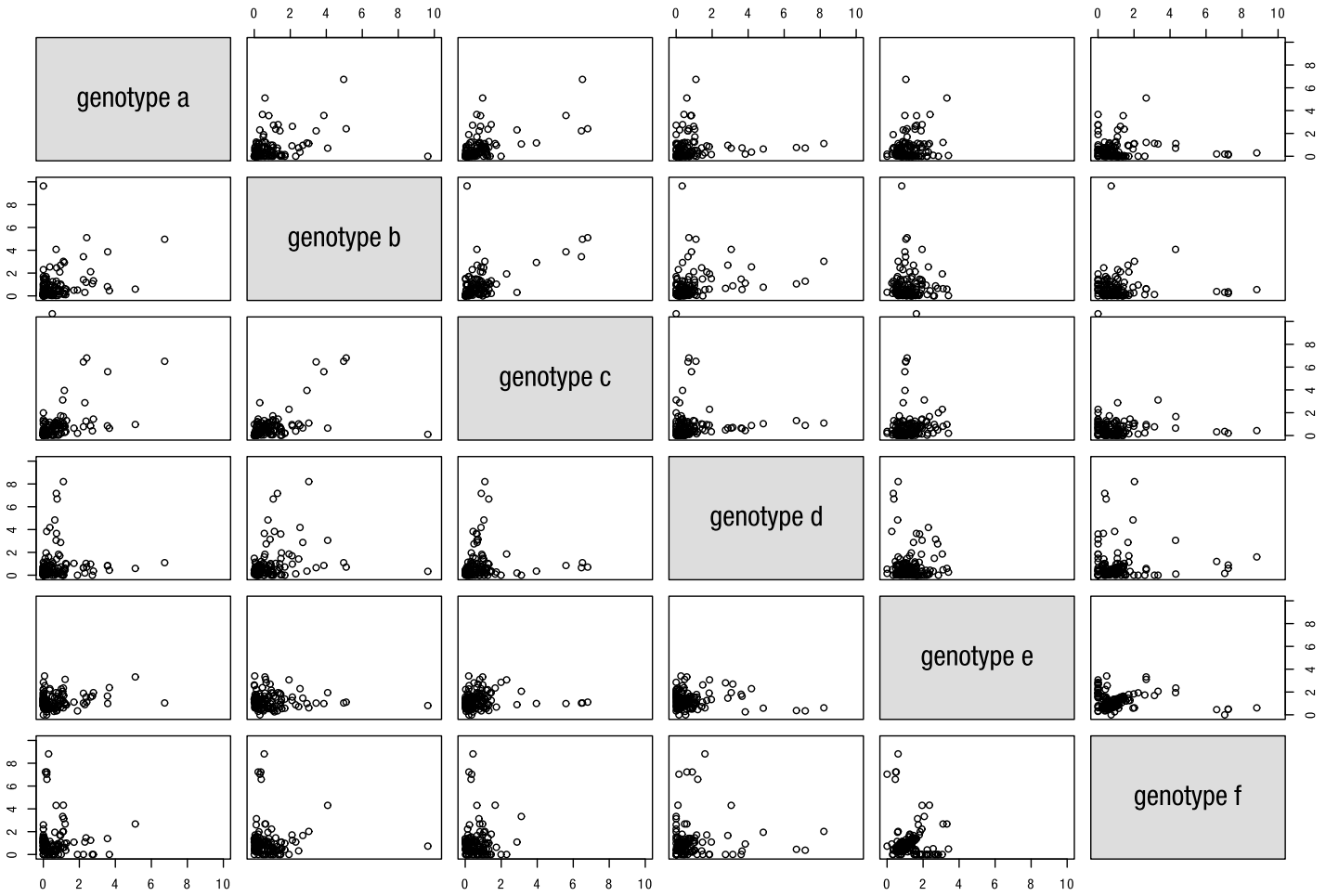

Figure 4.3: Matrix plot comparison of K_a/K_s ratios of HBx between genotypes A-F. K_a/K_s values were calculated using a sliding window of 15 nucleotides and a step of 3. The axes have a maximum arbitrary value of 10 and plots on opposite sides of the grey boxes mirror each other. A value less than one indicates negative selection, a value equal to one indicates neutral selection and a value greater than one indicates positive selection at a particular site.

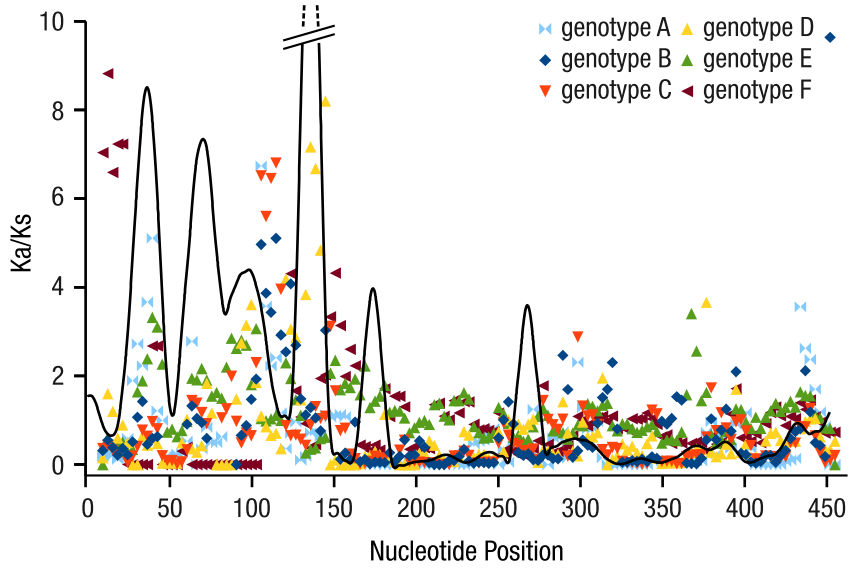

Figure 4.4: K_a/K_s values calculated over the length of HBx plotted against nucleotide position for genotypes A-F. The axes have a maximum arbitrary value of 10. Genotypes are represented by colour as follows: A - light blue, B - dark blue, C - orange, D - yellow, E - green, F - purple. A 4x moving average of K_a/K_s was calculated and plotted to assist with the interpretation of the overall trend (black line).

4.3.3 Sites Under Positive Selection

To identify sites within HBx subjected to positive selection (also known as diversifying or adaptive selection) several site-specific models within the PAML software were applied[215,216] to determine which model fitted the data. (1) Model 0 assumes a constant rate among all nucleotide sites and branches within the dataset. Genotypes A-E showed very similar dN/dS ratios (ω, ratio of non-synonymous over synonymous substitutions) with a mean of 0.673. Transition transversion (ts/tv) ratios were also similar with a mean of 2.546 (Table 4.1). (2) Model 1 suggests that HBx genotypes A-E are under negative selection (purifying or stabilising) such that the removal of these sites are deleterious (ω <0.15).

(3) Model 2 differs from Model 1 by incorporating an extra parameter for sites with an ω >1. Using this model, 64 % of HBx sequences in genotypes A-E are under negative selection, while 5-9 % are under positive selection. All PAML models are based on maximum log likelihood estimates (lnL) and a comparison of these values using a Likelihood ratio test (LRT) between Models 1 and 2 provide additional support for the data described above (Table 4.2). Furthermore, Model 2 provided a better fit of the data as compared to Model 1 (P>0.0001).

The nested site Models 7 and 8 employ 10 and 11 class site parameters respectively for the accumulation of ω values at sites along the gene. Model 8, parameter 11 shows dN/dS and LRT values similar to those presented in Model 2 for genotypes A and C-E (ω >1). Using this model

genotype B had a dN/dS ratio half that predicted by Models 1 and 2.

Models 7 and 8 also describe the shape of the β-distribution (dN/dS ratio as a function of the proportion of sites with a certain ratio). These are given as parameters p and q and extreme values indicate that either the proportion of sites are highly conserved or nearly neutral. Genotypes A-E and G have similar p and q values indicating that their β-distributions are U-shaped (see Table 4.1). Genotypes F and H have unimodal β-distributions for Model 8 as the p and q values were both above 1.

The results for genotype F, G and H were included in Tables 4.1 and 4.3 and Fig 4.5, but the low number of sequences analysed led to insufficient diversity within the dataset and thus these results were not discussed.

Table 4.1: Summary of Log-likelihood values (lnL) and parameter estimates for models applied to HBx across all genotypes of HBV using PAML Models 0, 1, 2, 7 and 8. Transition-transversion (ts/tv) ratios are represented for Model 0 only. $d_N/d_S(p_n)$ represents the proportion of data with a certain ω value. p and q denote the shape of the β-distribution of Models 7 and 8 only. For example, genotype A Model 1, $d_N/d_S(p_0)$ shows that (0.617) or 61.7 % of the data has an ω of 0.052.

	genotype A	genotype B	genotype C	genotype D	genotype E	genotype F	genotype G	genotype H
Model 0								
ts/tv	2.241	2.459	2.686	2.708	2.637	2.895	1.425	3.175
lnL	-2735.253	-3990.912	-3914.499	-3905.945	-1925.287	-1541.530	-876.802	-911.453
d_N/d_S	0.611	0.699	0.726	0.601	0.729	0.976	0.857	0.523
Model 1								
lnL	-2659.660	-3894.022	-3827.891	-3790.433	-1880.208	-1526.890	-876.727	-910.611
$d_N/d_S(p_0)$	0.052 (0.663)	0.100(0.636)	0.142(0.700)	0.101(0.730)	0.052(0.714)	0.000(0.490)	0.000(0.159)	0.088(0.554)
$d_N/d_S(p_1)$	1.000 (0.337)	1.000(0.364)	1.000(0.290)	1.000(0.270)	1.000(0.286)	1.000(0.511)	1.000(0.841)	1.000(0.446)
Model 2								
lnL	-2629.314	-3841.504	-3755.892	-3742.459	-1840.239	-1513.184	-876.710	-910.108
$d_N/d_S(p_0)$	0.052(0.617)	0.103(0.607)	0.162(0.647)	0.129(0.700)	0.062(0.650)	0.250(0.786)	0.000(0.239)	0.389(0.965)
$d_N/d_S(p_1)$	1.000(0.292)	1.000(0.378)	1.000(0.318)	1.000(0.207)	1.000(0.293)	1.000(0.000)	1.000(0.000)	1.000(0.000)
$d_N/d_S(p_2)$	4.304(0.091)	8.710(0.015)	6.820(0.035)	3.751(0.093)	7.539(0.058)	3.810(0.214)	1.150(0.760)	4.410(0.035)
Model 7								
lnL	-2660.299	-3900.615	-3839.132	-3806.842	-1882.259	-1526.487	-876.737	-910.628
p	0.017	0.228	0.329	0.202	0.013	0.005	0.063	0.022
q	0.021	0.296	0.459	0.361	0.025	0.005	0.012	0.020
Model 8								
lnL	-2629.599	-3849.545	-3761.387	-3744.251	-1840.348	-1513.191	-876.711	-910.110
p	0.173	0.276	0.400	0.402	0.098	33.194	0.005	63.198
q	0.384	0.371	0.489	0.976	0.181	99	0.008	99.000
p0	0.903	0.93	0.963	0.89789	0.942	0.786	0.392	0.035
$d_N/d_S(p_{11})$	4.013(0.097)	4.188(0.07)	6.727(0.037)	3.555(0.102)	7.511(0.058)	3.812(0.214)	1.179(0.610)	4.413(0.035)

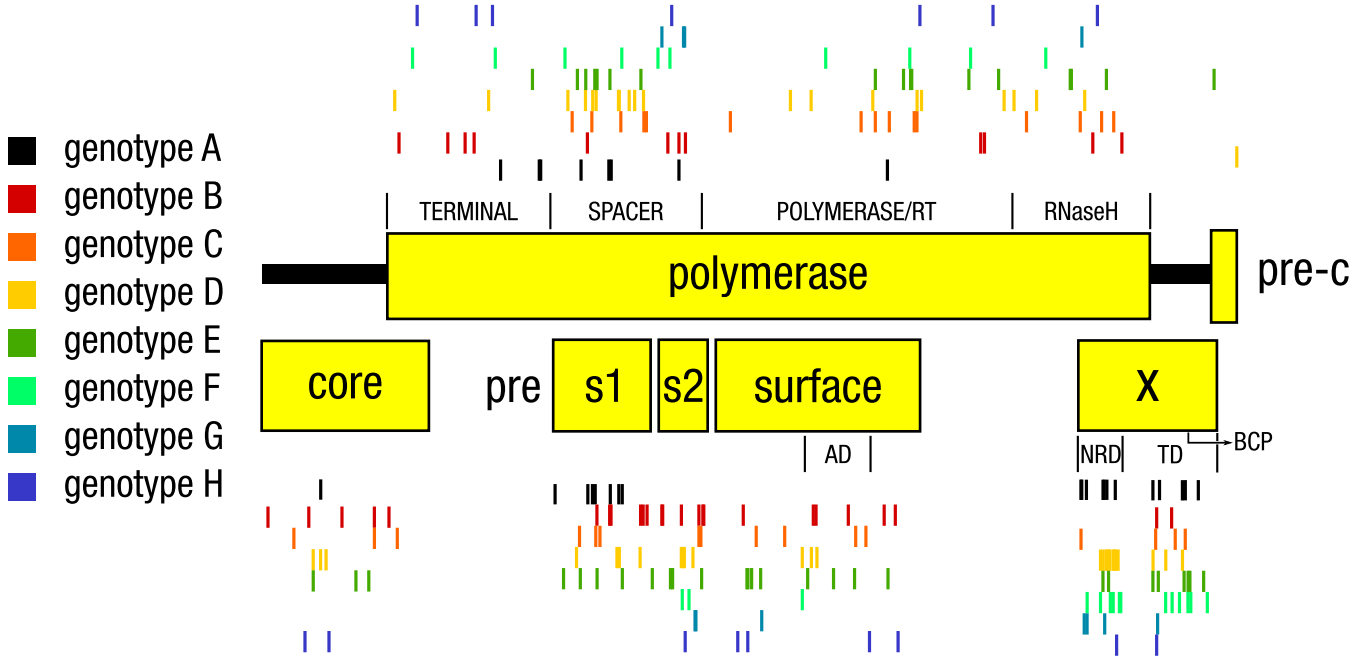

Figure 4.5: Sites within HBx under positive selection as determined by Model 8 of PAML software. All genes within HBV are shown in relation to each other as yellow blocks. The sites under positive selection are indicated as vertical lines according to different colours for each genotype. Several sites within the sequence of HBx are under positive selection both within the NRD and TD. There is a notable lack of sites under positive selection between the AD and NRE (seen as a gap). AD - *a* determinant, NRD - negative regulation domain, TD - transactivation domain, BCP - basic core promoter.

Table 4.2: Likelihood ratio tests (LRT) comparing the lnL values of PAML Models 1 versus 2 and Models 7 versus 8. Genotypes A-F were evaluated at confidence levels of P = 0.001, whereas genotypes G and H were evaluated at a confidence levels of P = 0.05 (as explained in the text). As the LRT values were greater than the critical value, it was concluded that Models 2 and 8 showed the best fit.

	genotype A	genotype B	genotype C	genotype D	genotype E	genotype F	genotype G	genotype H
LRT (2vs1)	60.692	105.036	143.998	95.948	79.938	27.412	0.034	1.006
LRT (8vs7)	61.400	102.140	155.490	125.182	83.822	26.592	0.052	1.036
Critical Value	13.815	13.815	13.815	13.815	13.815	13.815	5.991	5.991
P Value	0.001	0.001	0.001	0.001	0.001	0.001	0.05	0.05

Based on the LRT values (Table 4.2) Models 2 and 8 showed the best fits of the data and thus the sites identified as being under positive selection by these models were subjected to Bayes Empirical Bayes statistical analysis (BEB). The BEB output determines which amino acid sites were under positive selection and are displayed graphically in Fig 4.5 (As Models 2 and 8 identified the same amino acids, only one set of data was plotted). Sites under positive selection with HBx seemed to cluster in two distinct regions. The region between amino acids 1-50 in HBx, (predicted in Chapter 2 to show structural disorder Fig 2.14), showed a strong localisation of positive or diversifying sites under selection. The trans-activation domain (defined as amino acids 51-154) also showed numerous sites under positive selection with a notable region (amino acids 50-70) showing an absence of sites under positive selection. Taken together these data suggest that distinct regions within HBx are under positive selection across all genotypes tested although the reasons for this remain speculative.

Taken together these data suggest that distinct regions within HBx are under positive selection across all genotypes tested although the reasons for this remain speculative.

4.3.4 Branched Sites Selection

To determine if there were genotype-specific sites under positive selection within HBx, two branched site models (A and B) were applied. The data are represented in Table 4.3 and Fig 4.6. Model A suggested that 63-78 % of sites within HBx across all lineages are under negative selection ($\hat{\omega}_0$ $<$1), whereas 17-23 % of sites are nearly neutral ($\hat{\omega}_1$ = 1). The number of sites under positive selection varies widely between 0.07 % in genotype F and 11 % in genotype E. Genotype G shows a large number of sites under selection (20 %), but could be a consequence of the low number of sequences analysed and thus the result of a sequence bias. Genotypes B, D, F and H show a $\hat{\omega}_2$ = 999 which is the upper bound set for $\hat{\omega}_2$ and indicates infinity. This means that there are one or

more non-synonymous substitutions and no synonymous substitutions in the branch.

An infinity value makes the data difficult to analyse and thus an LRT should ideally be performed (using an lnL between branched site Model A and site Model 1, and branched site Model B and site Model 3).

However the datasets between the branched sites models and the sites specific models were different. The former required the use of close to 1000 sequences from different genotypes[217], whereas the site-specific models were applied to HBx sequences of only one genotype at a time. Nonetheless the data indicates that there are sites within each genotype of HBx that are experiencing positive selection.

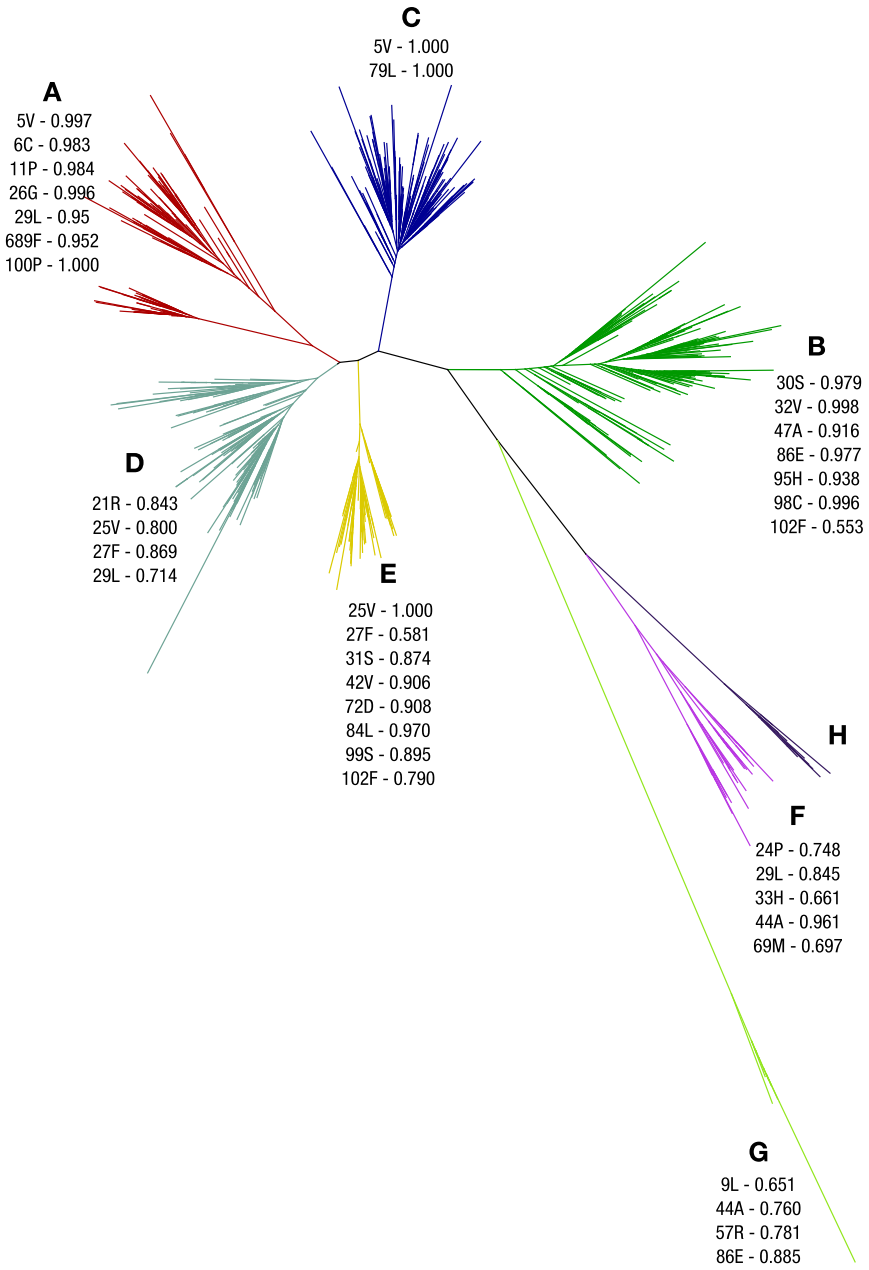

Figure 4.6: Phylogenetic tree of HBx sequences that were used to determine sites under positive selection. Sequences within each genotype are clustered in different branches based on sequence relatedness. The tree was constructed using BioNJ with 100 bootstraps. Each genotype is represented in a different colour. Specific amino acids are indicated with their probabilities of being under positive selection.

Table 4.3: Branched site model parameter estimates for HBx sequences across genotypes A-H. In Model A, $\hat{\omega}_0$ and $\hat{\omega}_1$ are fixed, whereas in model B they are free to vary.

Genotype	*l*	Estimates of Parameters	Positively Selected Sites
genotype A			
Model A	-8754.12	$\hat{p}_0$ = 0.714, $\hat{p}_1$ = 0.202, ($\hat{p}_2$ + $\hat{p}_3$ = 0.084), $\hat{\omega}_2$ = 4.736	5V 6C 11P 26G 29L 89F 100P (>0.95)
Model B	-8716.36	$\hat{p}_0$ = 0.846, $\hat{p}_1$ = 0.099, ($\hat{p}_2$ + $\hat{p}_3$ = 0.055) $\hat{\omega}_0$=0.310, $\hat{\omega}_1$=2.278, $\hat{\omega}_2$=5.994	
genotype B			
Model A	-8785.33	$\hat{p}_0$ = 0.781, $\hat{p}_1$ = 0.218, ($\hat{p}_2$ + $\hat{p}_3$ = 0.001), $\hat{\omega}_2$ = 999	30S 32V 47A 86E 95H 98C (>0.92)
Model B	-8729.168	$\hat{p}_0$ = 0.417, $\hat{p}_1$ = 0.071, ($\hat{p}_2$ + $\hat{p}_3$ = 0.513) $\hat{\omega}_0$=0.335, $\hat{\omega}_1$=2.415, $\hat{\omega}_2$=0.137	
genotype C			
Model A	-8737.935	$\hat{p}_0$ = 0.747, $\hat{p}_1$ = 0.232, ($\hat{p}_2$ + $\hat{p}_3$ = 0.021), $\hat{\omega}_2$ = 9.608	5V 79L (=1)
Model B	-8750.280	$\hat{p}_0$ = 0.850, $\hat{p}_1$ = 0.150, ($\hat{p}_2$ + $\hat{p}_3$ = 0.000) $\hat{\omega}_0$=0.295, $\hat{\omega}_1$ = 2.034, $\hat{\omega}_2$ = 3.071	
genotype D			
Model A	-8785.544	$\hat{p}_0$ = 0.780, $\hat{p}_1$ = 0.218, ($\hat{p}_2$ + $\hat{p}_3$ = 0.002), $\hat{\omega}_2$ = 999	21R 25V 27F 29L (>0.71)
Model B	-8737.240	$\hat{p}_0$ = 0.627, $\hat{p}_1$ = 0.110, ($\hat{p}_2$ + $\hat{p}_3$ = 0.263) $\hat{\omega}_0$=0.313, $\hat{\omega}_1$=2.087, $\hat{\omega}_2$= 0.00	
genotype E			
Model A	-8771.196	$\hat{p}_0$ = 0.688, $\hat{p}_1$ = 0.203, ($\hat{p}_2$ + $\hat{p}_3$ = 0.110), $\hat{\omega}_2$ = 5.023	25V 31S 42V 72D 84L 99S 102F (>0.89)
Model B	-8748.111	$\hat{p}_0$ = 0.850, $\hat{p}_1$ = 0.150, ($\hat{p}_2$ + $\hat{p}_3$ = 0.00) $\hat{\omega}_0$=0.294, $\hat{\omega}_1$=2.025, $\hat{\omega}_2$ = 0.00	
genotype F			
Model A	-8786.054	$\hat{p}_0$ = 0.781, $\hat{p}_1$ = 0.218, ($\hat{p}_2$ + $\hat{p}_3$ = 0.0007), $\hat{\omega}_2$ = 999	24P 29L 33H 44A 69M (>0.66)
Model B	-8748.521	$\hat{p}_0$ = 0.850, $\hat{p}_1$ = 0.150, ($\hat{p}_2$ + $\hat{p}_3$ = 0.00) $\hat{\omega}_0$ = 0.294, $\hat{\omega}_1$=2.025, $\hat{\omega}_2$=46.987	
genotype G			
Model A	-8781.898	$\hat{p}_0$ = 0.631, $\hat{p}_1$ = 0.169, ($\hat{p}_2$ + $\hat{p}_3$ = 0.200), $\hat{\omega}_2$ = 1.986	9L 44A 57R 86E (>0.65)
Model B	8743.369	$\hat{p}_0$ = 0.598, $\hat{p}_1$ = 0.105, ($\hat{p}_2$ + $\hat{p}_3$ = 0.297) $\hat{\omega}_0$=0.285, $\hat{\omega}_1$=2.035, $\hat{\omega}_2$=1.731	
genotype H			
Model A	-8793.564	$\hat{p}_0$ = 0.782, $\hat{p}_1$ = 0.218, ($\hat{p}_2$ + $\hat{p}_3$ = 0.00), $\hat{\omega}_2$ = 999	No positive sites
Model B	-8747.710	$\hat{p}_0$ = 0.625, $\hat{p}_1$ = 0.110, ($\hat{p}_2$ + $\hat{p}_3$ = 0.265) $\hat{\omega}_0$=0.294, $\hat{\omega}_1$=2.037, $\hat{\omega}_2$=0.246	

4.3.5 Posterior Probability Estimates of Omega

The parameters estimates described above focussed on specific sites within HBx under positive selection. The posterior probability estimate from PAML was used to show the proportion of negative (ω <1), neutral (ω = 1) or positive selection (ω >1) at each nucleotide across HBx. Thus this final estimate provided a more complete depiction of selection across HBx in different subtypes.

Fig 4.7-A shows the average selection probabilities across all genotypes. Fig 4.7-B and Fig 4.7-C display specific selection probabilities for genotype A and C respectively. Overall the average data showed that the disordered regions (25-55) as well as a few clusters in the transactivation domain have a higher probability of being under neutral or positive selection. This is in contrast to the data for genotypes A and C, where numerous sites have a higher probability of being under positive selection. Comparisons between the genotypes-specific and averaged datasets highlights the inherent problem with averaging data.

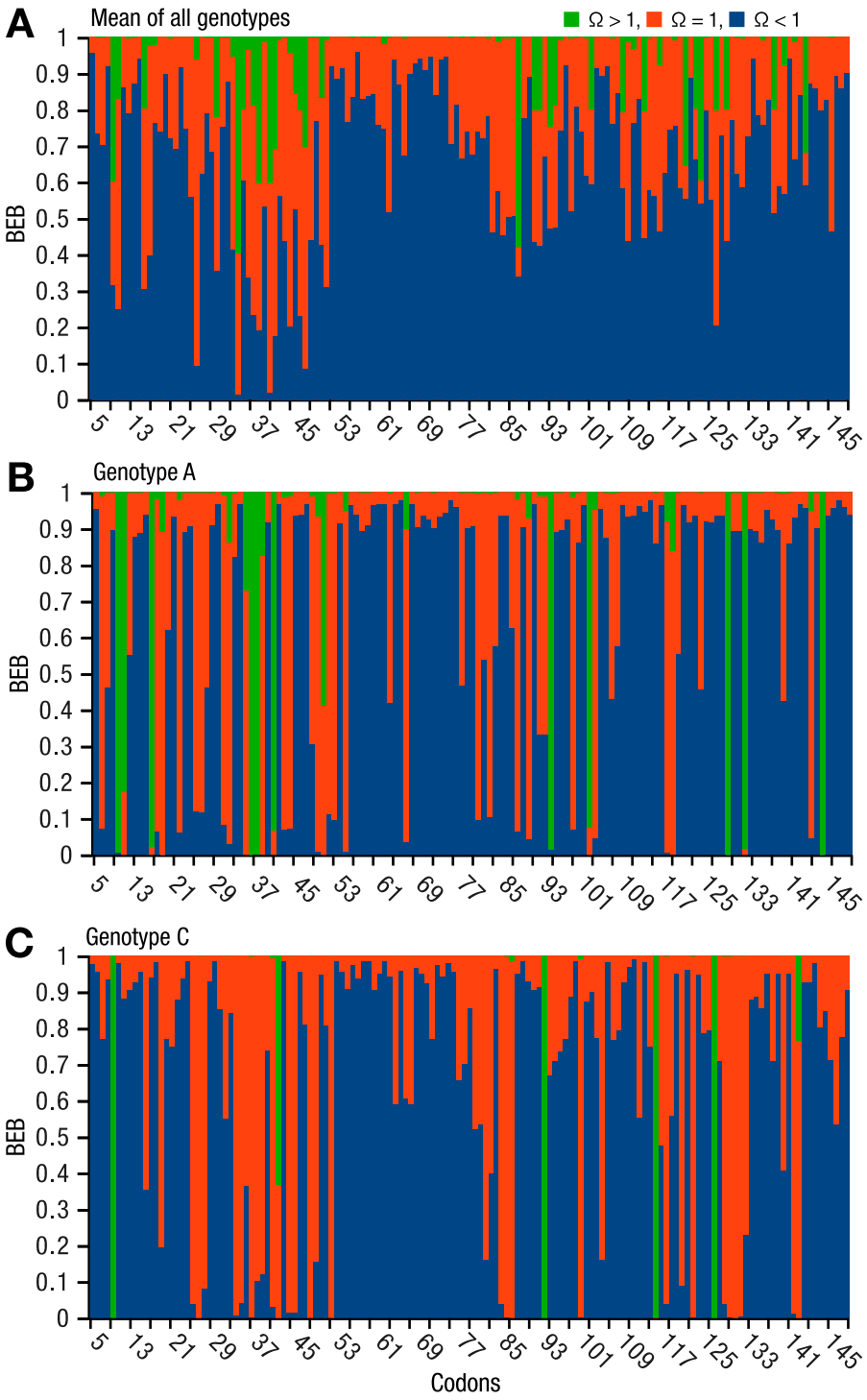

Figure 4.7: Posterior probabilities across all nucleotides within HBx displayed as Bayes Empirical Bayes values per codon. A - mean data across all genotypes, B - data for genotype A, C - data for genotype C. Sites under positive selection are displayed in green, sites under neutral selection are displayed in orange and sites under negative selection are displayed as blue.

4.4 Discussion

The HBV genome provides an interesting model with which to study the evolution of overlapping sequences. Studies have previously focussed on the *surface* and *polymerase* genes within HBV, as *surface* is entirely encoded within the *polymerase* gene and has a p3/s1 orientation (i.e. the third nucleotide of each *polymerase* coding triplet corresponds with the first nucleotide of each *surface* triplet, see Fig 4.1). However, the evolution of the HBx gene has not previously been described. As *hbx* overlaps with *polymerase* on its N-terminus and *pre-core* on its C-terminus, as well as having a non-overlapping region in the middle of its sequence, it was hypothesised that these three segments of *hbx* have different evolutionary constraints.

Entropy Initially, a cumulative entropy value was calculated for each of the three nucleotide positions of *hbx* coding triplets. This calculation is simply a measure of the sequence variability in an alignment at a particular position. The higher the variability, the higher the entropy value; the lower the entropy value the more certainty in the prediction of which nucleotide or amino acid is present at a particular position. Substitutions in the 3 different nucleotide positions will have different outcomes for the sequence concerned, and in overlapping genes the overprinted gene would importantly, be affected as well.

Evolution in the segment of *hbx* that overlaps with the *polymerase* gene, (encompassing 54 % of HBx protein length or amino acids 1-83) occurs mainly as a result of x1/p3 substitutions. This mechanism was evident in all the genotypes and importantly was similar to that reported by Zaaijer *et al*, where evolution of the *surface* region was found to occur primarily at s2/p3 sites. This result means that a nucleotide substitution in the x1 position, which would be non-synonymous in HBx is likely to be synonymous in P (p3). Additionally, mutations in the x2 and x3 positions were shown to be less frequent for this domain of HBx (Fig 4.2).

The middle section of *hbx*, which does not overlap with other HBV genes showed a different pattern of entropy. Evolution still occurred primarily via substitutions at x1, but there was a marked increase in substitutions at x2 positions. This is likely to be a result of the relaxation of substitution constraints compared to segment one which were dictated by the *polymerase* overlap. Substitutions at x3 sites were still as rare as in the first segment of *hbx*.

The last segment of *hbx* which overlaps with the *pre-core* section of the *core* gene did not show substantial entropy as shown by the low entropy values for these six amino acids. Nonetheless, evolution in this section was shown to occur primarily in the x2/c3 and x3/c1 positions and is thus consistent with reports published for the *surface* and *polymerase* overlapping domains[200].

K_a/K_s Analyses After determining at which codon positions substitutions were occurring, it was important to identify whether the evolutionary pressures were under positive selection or whether

they were merely polymorphisms.

The K_a/K_s ratio, which can be used to identify regions under positive selection was calculated first. Positive selection was suggested to occur in the following three sections of *hbx*: in genotypes A, E and F, there was a distinct spike in positive selection between amino acids 0-40; a cluster of positive selection within all the genotypes, between amino acids 80-160 (with values between two and eight) was observed; the trans-activation domain showed 3 spikes around nucleotides 300, 375 and 440 although values were between two and three. A matrix plot of the sliding window K_a/K_s values indicated that although most of the amino acids are under negative selection and thus clustered below one, there are distinct amino acid positions within certain genotypes that are experiencing positive selection, and are unique to that genotype. This may mean that HBx from the different genotypes is experiencing selection pressure unique to that genotype such as geographic separation, differences in ethnicity and healthcare disparities.

There are several sites in genotype A under positive selection, which are also under positive selection in genotypes B, C and E. A comparison of genotypes A, B and C with genotype E shows that there are sites under positive selection unique to the latter that are not present in the former three. A clearer illustration that some of the HBV genotypes are under different selective pressures is shown when comparing genotypes A and D and C and D, where sites showing K_a/K_s ratios substantially >1 in one genotype, are under negative or neutral selection in another genotype.

Even though the K_a/K_s ratio is a fair indicator of positive selection at the sequence level, it is unable to measure evolutionary changes brought about by changes to regulatory regions of genes, making it suitable for protein coding genes only. K_a/K_s also has limited power in determining selection in closely related lineages (such as the sequences used in this study) and cannot consider balancing selection where a polymorphism is maintained in a population. Many approaches average the K_a/K_s across the gene however, if a mean $K_a/K_s > 1$ is calculated, regions within the gene may still be experiencing negative or neutral selection (for example). Thus it is important not to average when calculating the K_a/K_s values for a gene. Furthermore, as it is more common for transitions (T-C and A-G) to be favoured over transversions (result of the bias in the frequency at which various nucleotides are swopped[218]), models need to account for non-homogenous rates of exchange[219]. The Nei and Gojobori method used in this chapter overestimates non-synonymous substitutions and underestimates synonymous substitutions. Despite this, the software package provided a less complicated result output as compared to other packages. In addition this method included sliding window analysis which generated a K_a/K_s for HBx without averaging.

The analyses presented in this chapter do not take in account *when* sequences were initially described. As it can take several generations for deleterious mutations to be removed from a population, it would be useful to include time as a component of the analysis. This would be particularly challenging as even though there are 350 million chronically infected people with the virus, there are

just over 2 000 full genomic sequences of HBV in public databases. Many of the entries are not well annotated which leaves an even smaller dataset with which to work. Despite these limitations, the results generated by the K_a/K_s sliding window analyses gave broad (gross) information regarding regions undergoing positive selection within HBx.

PAML Site Selection Analyses To further examine potential sites within HBx under positive selection, a comprehensive set of analyses were applied using other rigorous models. The dataset of re-genotyped and re-annotated HBV genome sequences was analysed by various models using the PAML software suite[208]. As displayed graphically in Fig 4.5, there are several sites within HBV and HBx that are specifically under positive selection. Amino acids and probabilities of positive selection are presented in Appendices A.1, A.2, A.3 and A.4 and only data for HBx is discussed further.

HBx shows an interesting partition of negative selection in the segment of the gene which forms the boundary between the negative regulatory domain and the trans-activation domain (nucleotides 200-260 in Fig 4.4). The negative selection observed in this region suggests that amino acid changes in this region are not tolerated thus highlighting the importance of this sequence within HBx. This approach provided a more accurate description of where mutations in HBx are likely to occur, and importantly, where in the gene they are likely to become fixed in the population. As with the calculation of the K_a/K_s, it would also be useful to include the effect of time, as in the analysis of closely related sequences, it is uncertain if these sites (calculated using PAML) represent fixation events in different genotypes, or polymorphisms that have segregated within a population[220].

PAML Branched Sites Selection Assuming that two lineages of HBx are evolving under purifying selection, and that the lineages are periodically subjected to bursts of adaptive change, a comparison of two sequences is unlikely to show a d_N/d_S ratio significantly greater than 1. Thus positive selection in one lineage is obscured by the other lineages as a result of the methods used to analyse the data. Therefore methods have been developed to determine whether there are different sites under positive selection, in different lineages of a phylogeny. PAML contains such methods which are based on a maximum likelihood approach and the models were applied to a phylogenetic tree constructed using all the HBx sequences used in this chapter.

The application of PAML branched site tests[211,221], where each genotype (the foreground) is compared to all the other genotypes (the background), showed that there are sites in almost all HBV genotypes that are under positive selection (Fig 4.6 and Table 4.3). This reflects what was observed in both the K_a/K_s ratio calculations as well as the site models (Table 4.1).

Generally models used in the study of the evolution of sequences are designed for certain categories of sequences, for example divergent as opposed to closely related sequences. Unfortunately, PAML suffers from limitations in the analysis of closely related sequences. The calculation of ω

(dN/dS) by PAML explicitly ignores polymorphisms that segregate within a population and instead represents each divergent species as a single sequence. Additionally, the package does not consider mechanisms by which a substitution enters a population and eventually becomes fixed. As stated by Kryazhimskiy and Plotkin, these simplified assumptions are acceptable when analysing substitution rates between long divergent species[220]. Additionally, all amino acid changes are assumed to be advantageous which is unrealistic, and as substitutions of amino acids are known to correlate with chemical properties, it is unclear how this information could be incorporated into current models[222,223].

Despite the limitations of the PAML analysis presented in this work, the results do demonstrate that there is strong positive selection on the negative regulatory domain which was predicted to be disordered. This suggests that the segment evolved more recently in a *de novo*, over-printing fashion. The evolution of HBx presented in this chapter fits with the three scenarios proposed by Rogozin *et al* for the modes of evolution in overlapping genes[224]. Firstly, the C-terminus of an overlapping protein evolves neutrally or almost neutrally. Secondly, the new protein coding region undergoes positive selection of non-synonymous replacement substitutions, thus altering its physical properties quite rapidly. Thirdly, the terminal regions of the two overlapping genes experience different modes of evolution.

The selection factors responsible for differential evolution in *hbx* could be as a result of stochastic processes in host-pathogen interactions, especially as the different genotypes are geographically separated. Additional factors could include HBV vaccine coverage, the proportion of the host population infected with HBV, regional differences in health care provision, cultural practises related to exposure to bodily fluids, or concurrent infections with other pathogens or parasites. The establishment of causality for the differences in positively selected sites between genotypes is not possible, and is thus open to speculation.

Applications of Evolutionary Patterning Analysis Phylogenomic analysis involves the comparison of genes and gene products, generally in the context of a whole genome to determine conserved functional residues. Such an approach offers insights into the evolutionary process and is frequently used in the drug design process[225]. Current treatment regimens for chronic HBV infection rely on nucleotide and nucleoside mimics to inhibit viral replication. Therefore, although there are no peptide inhibitors available for the treatment of HBV, there exists a useful method of determining the best possible target sites that would be effective across multiple genotypes. The evaluation of a plot of cumulative probabilities in Fig 4.7 demonstrates quite clearly which sections of HBx are subjected to positive evolutionary pressures and those that are not. This evaluated output, which can be applied to any sequence of interest, is known as evolutionary patterning, and has been previously applied to *Plasmodium falciparum*[226]. Sequences under strong negative selection are unlikely to

undergo substitution and thus present attractive targets. Importantly, different evolutionary pressures in the different genotypes of HBV means that different sections of proteins are more or less likely to undergo mutation. This is illustrated in Fig 4.7. It is important to consider this information when designing chemotherapeutic agents, such that the rapid generation of drug resistant mutants can be mitigated against.

Overall the analysis of evolution across multiple related pathogen sequences such as those in the HBV family has not been exploited. The power of these approaches is under appreciated, particularly with regards to human pathogens that cause a high disease burden in Africa. Computational biology and evolutionary studies will most definitely play a more prominent role moving forward and should be more readily included in research approaches.

Chapter 5

General Discussion and Conclusions

HBV has been implicated in disease since 1947, yet despite discovery of the HBx protein in the early 1980s, controversy remains as to its function in both the viral lifecycle and progression of HBV-associated HCC. Structural studies of other pathogen-derived proteins have been of benefit to targeted drug development in diseases such as malaria (caused by *Plasmodium falciparum*). Conversely HBx structural studies have been hampered by expression difficulties and as a result no X-ray or NMR data are available for HBx. Furthermore, the role that HBx plays in liver disease has been negatively affected by reliance on transfection-based overexpression systems and physiologically irrelevant cellular models. These methodologies may have inadvertently introduced numerous artefactual findings into the field.

The work presented in this thesis adds to the growing body of knowledge and methods used to study the function/s of HBx, with particular emphasis on the generation of soluble protein within eukaryotic cells. In addition an extensive evolutionary analysis across hundreds of HBV sequences, and an in-depth exploration into disordered proteins and antibody problems was conducted. Taken together these aspects as presented in this thesis provide a solid argument to explain some of the controversies around HBx and deliver an enlightening explanation of the complex HBx protein.

5.1 Approaches used to Express and Purify HBx

The yeast *S. cerevisiae* is a well established system that has been used by researchers to express a multitude of proteins of biomedical and biotechnological importance.

The expression of HBx protein in *S. cerevisiae* was guided by previous expression studies of the other major proteins of HBV. The HBV vaccine, derived from the HBs protein, was first produced using *S. cerevisiae* in 1982, under the control of an alcohol dehydrogenase (ADH1) promoter[227]. The Core protein, responsible for the core particles, was produced four years later under the control of a yeast acid phosphatase (APase) promoter[228]. The Polymerase protein was also expressed in yeast for drug testing although not in macro quantities[229]. Therefore it was reasoned that *S.*

cerevisiae would be a good system in which to express HBx.

In Chapter 2, the expression studies using *S. cerevisiae* were not initially deemed to be successful as no HBx protein could be detected from any of the vectors used. Two additional mammalian expression vectors mirrored the results of the yeast vectors i.e. no protein was detected on Western blots. As full length mRNA was detected for all vectors, an HBx fusion protein was generated to try and visualise HBx directly within the cells. Fluorescence and immunofluorescence microscopy data suggested that antibody binding to HBx in Western blotting analysis was the key problem. Overall this suggested that the denaturation steps required during Western blot analysis were altering the antigenic sites on HBx and thus negatively affecting antibody binding site recognition. As the HIS tag was not successfully used nor detected, future work should include multiple affinity tags. A combination of small affinity tags such as 6xHIS, cmyc, haemagluttinin and FLAG would offer a much higher chance of detecting and purifying recombinant HBx. A detailed critical discussion of antibody problems is presented later in this chapter (section 5.1.1).

In parallel with the expression of HBx from mammalian vectors, the use of an MBP fusion tag vector was also explored. This allowed purification of HBx to a high degree as well as major improvements in solubility and stability of the protein as compared to non-fusion protein systems. The MBP-HBX construct was sub-cloned into SFV expression systems to allow for eukaryotic expression. SFV has been used to express a myriad of proteins, including the structural proteins from HBV[171], suggesting that it would be suitable for the expression of HBx. Disappointingly only the pSFV-b12A-ppt-MBP-HBx showed detectable intra- and extracellular HBx protein, albeit with the continued presence of degradation products. This suggested that the fusion protein was only partially stable in SFV systems, and thus the yield of HBx could be improved with further modifications to the vector. The scaling up of the SFV expression system was also met with technical hurdles.

SFV has been successfully used in many studies to express a variety of proteins. However, the limited success in expressing HBx using this system suggests that there are factors influencing either the intracellular stability of the SFV replicon or that certain proteins (i.e. HBx) are difficult to express using SFV. It is important to determine whether these factors are general (e.g. an interferon response that expels the SFV replicon) or specific, such as protein toxicity. A combination of the cellular toxicity associated with the overexpression of intrinsically disordered proteins (IDPs), the difficulty in generating high titre virus with pSFV-b12A-ppt-MBP-HBx and the unproductive efforts in selecting clones of the ncSFV-pac2A-MBP-HBx suggests that it was a combination of general and specific factors. It would thus be worthwhile to determine the interferon response of BHK-21 cells before investing significant resources and time in trying to express difficult-to-work-with proteins such as HBx. It would be worthwhile testing other cell lines for the ability to express this construct as well and/or developing interferon knockout cell lines. Close collaboration with the developers of the SFV vectors will help ensure that the large-scale expression of eukaroytic HBx is met with future

success. The minimal expression of soluble HBx remains a significant step in the right direction and represents a promising point from which further expression studies can continue so that structural determination of HBx can be achieved.

5.1.1 Problems with Antibodies

The work in Chapter 2 highlighted the limitations in relying on antibodies for detecting recombinant proteins. Despite antibody detection being a well characterised and successful technology, there are instances where false positives and false negatives occur as a result of the incomplete characterisation of the antibodies. The frequencies at which these false results occurs is not known, although The Human Protein Atlas, (an initiative which hopes to create and characterise polyclonal and monoclonal antibodies for every human protein), has noted that only 35 % of commercial antibodies that have been tested in their strict quality control pipeline actually detect antigens as specified by the manufacturers[157]. This suggests that false results are fairly common and have undoubtedly led to artefactual findings detrimental to proteomics and many other fields of research.

The protein expression challenges in this project were typified by such misleading antibody activity. In-house anti-HBx antibodies were developed with the assumption that the structural differences between prokaryotic and eukaryotic HBx were minimal thus allowing both forms of the protein to be recognised by the antibodies. This was unfortunately not the case and highlights the major problem in developing monoclonal antibodies against *E. coli* derived antigens without adequate characterisation of the binding capacity to eukaryote-derived protein. An assumption such as this is likely to be correct in many cases, but as HBx is known to have a low antigenicity, there should have been a larger focus on prokaryotic versus eukaryotic binding validation.

Literature suggests that binding events should be allowed to occur in solution to encourage interactions between native protein forms and the antibodies[161,162] ans this is recommended for all future attempts at creating anti-HBx antibodies. Additional specifications should include confirmation that the antibodies are able to bind to eukaryotic HBx protein, whether by Western blot or ELISA, and an explicit statement listing which binding capabilities have and have not been tested.

The default use of a multiple antigenic tag (6xHIS, HA, FLAG and/or c-myc) together with a solubility-enhancing fusion protein (SUMO, NusA or MBP) is likely to ensure that the large scale reproducible expression of soluble HBx will be achieved.

5.2 Instability, Disorder and Moonlighting in HBx

The evidence of disorder in the secondary structure of HBx was initially shown by Liu *et al* from results of CD-spectroscopy using soluble, pure, *E. coli* derived HBx[128]. As CD-spectroscopy cannot infer where the structural disorder lies in the protein backbone, various predictive models were ap-

plied to the HBx sequence with results showing the presence of a major disordered region between amino acids 25-52. These predictions are shown in Fig 2.14 and whilst all four models showed similar outputs, PRDOS (Fig 2.14-D) indicated that there may be additional regions of disorder outside of amino acids 25-52. Based on the presence of multiple disordered regions within the sequence, it is thus possible to presumptively classify HBx as an Intrinsically Disordered Protein. However as this prediction is only based on CD-spectroscopy and data modelling, structural determination should be used to verify the disorder within HBx as it has important implications for HBx-related research.

As IDPs generally have an elevated protease sensitivity and a lower half life[143] this offered an explanation as to the poor stability and short half-life of HBx. Furthermore these factors could also have contributed to the technical challenges in working with HBx as detailed in Chapter 2 and 3. Intriguingly the characteristics of IDPs also apply to a specific group of proteins known as ”moonlighting proteins”. Constance Jeffery coined the term for those proteins with the ability to perform more than one cellular role[230]. Examples include enzymes, chaperones, transcription factors, and proteins with additional functions. The characteristics defining moonlighting proteins include one of more of the following: 1) the capacity to bind to different partners, 2) the performance of different functions in various parts of the cell, 3) the ability to function differently inside the cell compared to outside, 4) a change in function as a result of a change in temperature, 5) a change in the redox state of the cell, 6) distinct activities in diverse cell types, 7) the effect of oligomerisation, i.e. the same protein can have one function in a monomeric state and another in a multimeric state, 8) the ability to elicit both inhibiting and activating effects on different partners or even the same molecular partner, which may also depend on the concentration of the substrate, ligand or co-factor available. It is important to note that these characteristics are not mutually exclusive since a combination might be required for a protein to function. For example, band 3 of the red blood cell membrane is an anion exchanger but also regulates the rate of glycolysis in red blood cells[231]. There does not seem to be any overlying similarity between moonlighting proteins other than the ability to perform two or more disparate functions. Four examples of proteins that fulfil these criteria are briefly discussed below, and a more comprehensive list may be found in (Jeffery 1999)[230] and more recently in (Jeffery 2009)[232]. 1) *E. coli* PutA displays proline hydrolase and pyrroline-5-carboxylate dehydrogenase activity when associated with the plasma membrane, but when the protein is in the cytoplasm it binds DNA and acts as a transcriptional repressor[233]. PutA changes function as a result of substrate concentration i.e. the protein binds to the membrane when cytoplasmic substrate concentrations of proline are high, but binds to *puta* control-region DNA when cytoplasmic substrate concentrations decrease. 2) Neurophilin is a cell surface receptor on endothelial cells that detects endothelial growth factor and provides an indication of when new blood cells are needed[234]. Contrastingly, in nerve axons it serves as a detector of semaphorin III which helps direct axon growth[235]. 3) Thymidine phospho-

rylase has a different function inside as compared to outside of a cell. In the cytoplasm it catalyses the dephosphorylation of thymidine and deoxyuridine to their respective analogs. However in the extracellular milieu, it acts as a platelet-derived endothelial growth factor which is a stimulant of chemotaxis and endothelial growth. 4) Glyceraldehyde-3-phosphate dehydrogenase is a glycolytic enzyme that in a tetrameric state converts glyceraldehyde-3-phosphate to 1,3-diphosphoglycerate, whereas in a monomeric state is a nuclear uracil-DNA glycosylase[236].

Tompa *et al* suggested that some moonlighting proteins could have disordered structures[143]. The template-induced folding mechanism that many IDPs use to interact with binding partners may mean that a single disordered region is able to adopt a variety of conformations when it binds to different proteins. Additionally, activation or inactivation may occur depending on the binding partner, and the IDP may bind to the same partner at different sites or in different conformations[143].

In addition, IDPs are tightly regulated from transcript synthesis to protein degradation and studies have shown that most IDPs are present at relatively low levels, for short periods of time as compared to structured proteins. The short half-life at the transcript and protein levels may facilitate quick adjustments of protein concentrations when environmental conditions or signal inputs change. As tight regulation of IDPs is a conserved mechanism in *S. cerevisiae*, *Schizosaccharomyces pombe* and humans, this suggests that selective forces operate on both IDPs and their transcripts to precisely regulate abundance and half-life[237]. This is likely similar in HBx regulation and there is growing evidence that this tight regulation is exploited by viruses to establish successful infections. It is possible that exploitation of IDP regulation occurs via the use of mimicry i.e. by viruses encoding host-like peptide motifs which are able to titrate host cell proteins and thus alter conditions to make infection more successful[238–242]. Related to this, several important viral studies have uncovered a novel class of compact non-globular protein interaction interfaces[243–246]. These are called short linear motifs, SLiMS, LMs or minimotifs and are ubiquitous in eukaryotic genomes as well[247]. The defining features of short linear motifs include the capacity to encode a functional interaction interface of only three to ten residues, an enrichment in the intrinsically disordered regions of proteins, the ability to function independently irrespective of their tertiary structure context, and the tendency to evolve *de novo*. Additionally, linear peptide motifs have been shown to be dosage sensitive, and are thus tightly controlled to prevent toxicity upon overexpression under physiological conditions[247]. Vavouri *et al* suggests that disordered regions are prone to making promiscuous molecular interactions via linear motifs when their concentration is increased, and this is the likely cause of toxicity[247].

Linear motif-mediated interactions are abundant in nature, and as the motif binding domains are present in many families of related proteins with similar binding sequence preferences, there is a strong possibility that they will bind to each other's targets if their concentrations are increased within the cell[248,249]. Concomitantly, scaffolding proteins are critical regulators of many key cell signalling pathways and exhibit one-to-many binding often mediated by short linear motifs[250]. As

scaffolding proteins play a crucial role in coordinating specific interactions, they are required to be present in cells at a defined concentration range to avoid molecular titration or cooperative binding. It has also been noted that several of these scaffolding proteins possess disordered domains[251].

The inherent characteristics of moonlighting IDPs and scaffolding proteins offer the clearest explanation of the functions of HBx, provide reasons for contradictions in the literature and elude to its role in the development of HCC. The disordered region within HBx may be a mechanism used by the protein to mediate its many functions. As the predicted disordered sequence is only 25 amino acids in length and is located within a region under positive selection (i.e. prone to mutations), it provides the protein with a dynamic binding site capable of interacting with multiple partners. The short predicted disordered region of HBx could encode multiple linear motifs (typically 3-10 amino acids long) enabling it to bind to multiple protein partners. This has been confirmed as HBx localised with a cytoplasmic complex containing MEKK1, SEK1, SAPK, and 14-3-3 proteins using antibody pulldown assays. Mutational analysis of the RXRXXpS phosphorylation motif (amino acids 26–33 of HBx, pS is phosphoserine) of HBx confirmed that this motif was needed to bind the 14-3-3 scaffold proteins, which was in turn necessary for the induction of SAPK/JNK activity[252]. Separate linear motif interactions include DDB1 binding (HBx amino acids 84-101)[253], SH3 binding domain PxxxP starting at amino acid 26[254] as well as NES binding motifs (amino acids 89-100)[255].

Whilst the presence of linear motifs in HBx have been partially explored, these experiments regarded the linear motifs as static structures without cognisance of protein disorder. The total number of potential interactions was not explored. Furthermore the use of overexpression in some of the studies raises doubt as to the accuracy of the findings. Overexpression constructs specifically generate protein levels that far exceed physiologically normal concentrations thus encouraging non-specific interactions. Combined with the promiscuity of disordered proteins and the presence of linear motifs, overexpression studies of HBx have most likely generated spurious findings and incorrectly identified artefactual interactions. Nevertheless these studies have provided an initial indication that interactions between HBx and cellular proteins in part occur via linear motifs.

The disorder and linear motifs of HBx described above may also link to the moonlighting characteristics of this protein. As noted in the introduction of Chapter 2, conclusions regarding the sub-cellular localisation of HBx remain unresolved. Su *et al* established that anti-HBx immunocytochemical signals were only located in the cytoplasm of explanted liver cells[256], and this was also shown using liver cells stably transfected with HBV[82]. These bodies of work are important because naturally infected liver samples i.e. liver cells with no overexpression of HBx were used. Work by the same lab, but using cultured cells and CMV-driven HBx plasmids, established that HBx was present in both the cytoplasm and nucleus[84]. Conversely Haviv *et al*, also used cultured cells and overexpression plasmids and observed that HBx was primarily located within the nucleus[83]. Taken together these datasets highlight the discrepancies produced when disordered proteins such

as HBx are overexpressed in cell culture models, and argue strongly in favour of abandoning CMV-driven HBx expression in favour of tuneable or conditional expression systems (e.g. tetracycline or RNA interference based vectors[257]).

Overexpression studies with HBx have also identified a role for HBx in p53 modulation[258]. The frequency of p53 mutations in HBV-related HCC were lower than the frequency of p53 mutations in chemical carcinogen-related HCC[259–261]. The authors of this study suggested that this was as a result of HBx counteracting p53 function, thereby influencing cellular death and apoptosis. However as overexpression was used and some of the studies conducted in physiologically irrelevant mouse cell lines (NIH3T3) the data should be interpreted with caution. In particular if HBx is indeed an IDP the high number of interactions observed is not surprising and the relevance and accuracy thereof is questionable. Indeed in a study conducted with the effects of overexpression in mind and in relevant cellular models, Terradillos *et al* found that there was a direct, dose-dependent apoptotic effect of HBx in transiently transfected, liver-derived cell lines[262]. There was also evidence that the intentional overexpression of HBx induced apoptotic death. Contrary to the authors conclusions this indicates that these observations were due to the toxic effects of overexpressing a disordered protein combined with the disorder of p53 and the propensity of IDPs to form non-specific interactions at elevated levels. In summary, numerous observations suggest that HBx is an IDP that contains linear motifs capable of binding multiple protein partners, and when overexpressed, particularly in non-hepatic cells, artefactual interactions abound.

5.3 The Evolution of HBx

The compact nature of the HBV genome provides an interesting model for the study of overlapping genes and the efficient use of sequence space. This means that the HBV genome can encode fewer proteins overall but maintain multiple functions and interactions. HBx is particularly interesting as it overlaps with two other genes, contains intrinsic disorder and may function as a moonlighting protein. This is unusual as most overlapping proteins have only one overlapping partner. The measurement of entropy across any gene provides simplified information that is otherwise difficult to discern from substitutional models. In addition, measurements of entropy reduce the complexity of models, for example the higher the entropy for a given site within a sequence the greater the number of substitutions that occur at that site.

Entropy A comparison of the first segment of *hbx* which overlaps with *polymerase* (amino acids 1-83 Fig 4.2 Panel A), shows that the first codon position of *hbx* undergoes the most substitutional activity. This would result in a change within HBx without a change in Polymerase. However, the entropy in the x2 and x3 positions are generally equal as mutations here are likely to affect polymerase

and will thus be selected against. Taken as a whole, the pattern of reading frame substitutions is similar for most of the genotypes, however comparison between individual genotypes reveals many differences in the preference of one reading frame over another. For example certain genotypes (E, G and H) show higher activity in the x2 position while others (A and F) show higher activity in x3. The effect of the genotype-specific substitutions and preferences is suggestive of unique virus host interactions. Furthermore the effect the reading frame preferences have on the functioning of the numerous regulatory sequences contained within HBx are not dealt with in this work, and will provide an important angle for future analysis.

Positive Selection A broad overview of substitution activity using K_a/K_s estimation within *hbx* revealed that various sites were under positive selection, predominantly those predicted to be disordered. Further analysis using the PAML suite of evolutionary models confirmed these data and suggested that the negative regulatory and transactivation domains were also under positive selection. The data also showed a notable absence of sites under selection between these two domains. It is difficult to determine the relevance of these sites of positive and negative selection in the absence of complete structural data for HBx. However, analysis across all genotypes showed a similar trend of positive selection within the disordered region (amino acids 25-55). This suggests that the genotypes are experiencing similar evolutionary pressures within HBx which is interesting given that the genotypes of HBV are still largely geographically separated. Further analysis to determine if these positively selected sites are linked to particular linear motifs would be an interesting avenue of enquiry. This may also be a unique evolutionary strategy for using redundancy to allow multiple protein expression without an increase in the HBV genome size[263].

5.3.1 Evolutionary Patterning and Drug Targeting (Posterior Probability Estimates of Omega)

Evolutionary patterning is a tool that can guide the development of chemotherapeutic agents, small inhibitory molecules and peptides against protein targets of interest. The data generated using this technique allows one to determine whether amino acid sites are under positive or negative selection, and thus more or less likely to undergo substitution at some future point. This is especially applicable to viruses, as their mutation rates are several fold higher than humans or eukaryotes in general. Thus therapeutics designed against viral proteins are likely to have a shorter therapeutic lifespan.

Sites under negative selection present better targets due to the lower likelihood of resistance mutations developing. The region between the negative regulatory and transactivation domains of HBx is under negative selection. As there no sites under positive selection here, it suggests that this region is important for HBx function and thus substitutions within this region are selected against, thereby providing a potential drug target site (Fig 4.5). However without adequate structural data,

it is not possible to determine whether this would be viable or not. The disordered domain in HBx would not be a good target as it has numerous sites that are under positive selection and is thus more likely to undergo substitution.

As HBx interacts with numerous IDPs such as p53, it may be more beneficial to develop drugs against these targets or the regulatory enzymes that post-translationally modify IDPs instead[264]. IDPs are strictly regulated, but the action of viral infections on these proteins causes their dis-regulation, resulting in a disease phenotype. A viral protein such as HBx disregulates the human p53 IDP which is one of the steps involved in the development of HCC. The generation of small molecules aimed at restoring the levels of host IDPs (e.g. p53) via competitive binding to the affected IDP might be feasible course of treatment in the future. Importantly the propensity of IDPs to bind to multiple partners would probably be mirrored by the ability of small molecules to bind to multiple IDPs as well. Thus the spectre of off-target effects would need to be recognised and studied in detail. Two recent studies using disordered regions of cellular proteins as potential drug targets illustrate the potential of this approach: 1) the Myc transcriptional regulator is thought to regulate 15 % of genes, and is found in many cancers as a mutated species. Small molecules against Myc were found to shift the equilibrium of dimerisation of Myc with its binding partner Max (myc-associated factor X), thus counteracting the effects of Myc overexpression common to many cancers[265]. 2) BRD4 protein which associates with chromosomes during mitosis and to the RING3 serine/threonine kinase was targeted with a cell-permeable compound against its disordered histone tail in the bromodomain. This event was shown to prevent BRD4 from binding to the chromatin and thus established an anti-proliferative effect in human squamus epithelium cells[266]. Even though this strategy would require prior knowledge on the concentrations of cellular IDPs from patient to patient, it may be more successful long-term as the host target protein is less likely to undergo mutation.

5.4 Future Work and Recommendations

The work in this thesis focussed on the expression and purification of HBx and provides an excellent point from which further studies at improving the results presented, can be attempted. The difficulties associated with the low antigenicity of HBx should be noted, and only antibodies of high quality that have been rigorously tested should be used. Ideally, HBx should be expressed with a solubility-promoting tag (MBP, NusA or SUMO) that has another purification tag such as 6xHIS to permits rapid purification of the protein and maximum yield. Furthermore, considering the antigenic differences that were discovered between the *E. coli* and eukaryotic HBx protein, it is recommended that structural studies are pursued using eukaryotic HBx.

The presence of disorder in HBx is an important finding as IDPs are known to be difficult to crystallise and many structural biologists would consider it a waste of time and resources to do so.

Unless there are a few stable, low energy conformations that HBx can assume, the isomorphous lattice that is needed for X-ray studies will probably not form because of the structural interconversions of the protein[267]. It is thus suggested that structural studies of HBx protein should focus on a wide array of techniques including: small angle X-ray scattering, paramagnetic spin labelling in combination with ensemble molecular dynamics and single molecule fluorescence, high-speed atomic force microscopy, Raman optical activity and NMR (reviewed in[268]).

It is hoped that due to the characteristics of IDPs, the use of overexpression and irrelevant cellular models is rejected to prevent artefactual research findings from continuing to cloud the functions of HBx and its contribution to the development of HBV-associated HCC. The use of inducible and tunable expression systems but more importantly natural infection conditions is encouraged to ensure that the levels of HBx, and thus the extent of its interactions with cellular targets is physiologically relevant. The lack of availability of cell culture and small animal models of HBV infection has been an important limiting factor. Greater resources should be placed on elucidating the cellular receptor used by HBV to infect cells. Knowledge of the receptor will aid the development of cell culture and small animal models.

The evolutionary analysis of HBx provided important insights into the sequence plasticity of *hbx* and showed that the main disordered region is under strong positive selection. Whether this is a result of its *de novo* creation as discussed in Chapter 4, or interactions with the human immune system is a topic for future work. It is important to view bioinformatic analysis of *hbx* and protein function in general as a tool to improve the design and planning of wet lab experiments. For example, work in this thesis suggests that the development of inhibitors against HBx should take into account the presence of disorder and the propensity of substitutions to occur in these regions. It is suggested that inhibitor development focuses on cellular partners that HBx is known to interact with under normal physiological conditions.

This work will hopefully result in a re-evaluation of the methodologies used to gather data on the functions of HBx and other difficult to express proteins. Consideration of the role of IDPs in viral infection and cancer development will accelerate the description of the role that HBx plays in the development of HCC and will be useful in the design of inhibitors against HBx ligands. The high number of publications involving HBx conveys a sense of importance to this protein which may not a true indication of its biological activity. Upon personal reflection of the first papers linking HBx and HCC development, I concluded that the field was hungry for a single protein that could be responsible, and HBx became the *preferiti*. Two decades have followed with little change to the original methodologies and dogma. Consequently much artefactual data has been generated. Through rigorous analysis of all data with particular emphasis on negative results and by not compromising on methodologies used in this study of HBx, we have illuminated the need for realignment within the HBx field. It is hoped that this work will contribute to the understanding the role IDPs play in the

development of cancer. IDPs and linear motifs may provide a novel avenue for the development of anti-cancer drugs, especially when genes are over-expressed. Furthermore viral exploitation of host IDPs via linear motifs to establish infection provides targets to develop antiviral therapies. Hopefully another decade will not lapse before these suggestions are realised.

Appendices

Appendix A

Supplementary Data

A.1 Supplementary Data

A.1.1 Testing for overlapping antigenic site reactivity of anti-HBx antibodies

Antibodies that had been generated by Dr Wolfgang Prinz against a urea denatured 6xHIS purified, bacterial expressed HBx antigen were tested to make sure the antigenic sites did not overlap. It was simply determined whether the incubation of two different antibodies produced a doubling of the output reading (no overlapping antigenicity), or a change in the output reading (overlapping antigenicity). This data is not shown.

A.1.2 Immunofluorescence Microscopy Control Images

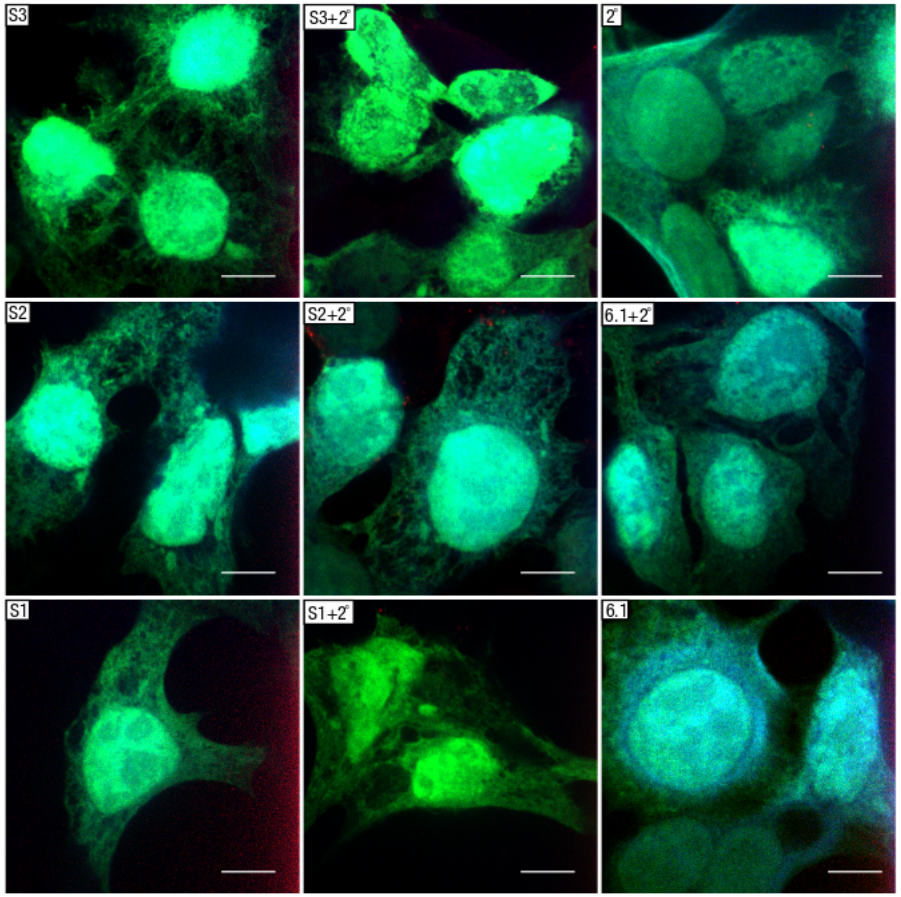

Figure A.1: Immunofluorescence microscopy images of HBx staining in HEK293 cells expressing the HBx negative, pCI-Neo-GFP vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in any samples. Magnification is 100x and scale bar = 5μm.

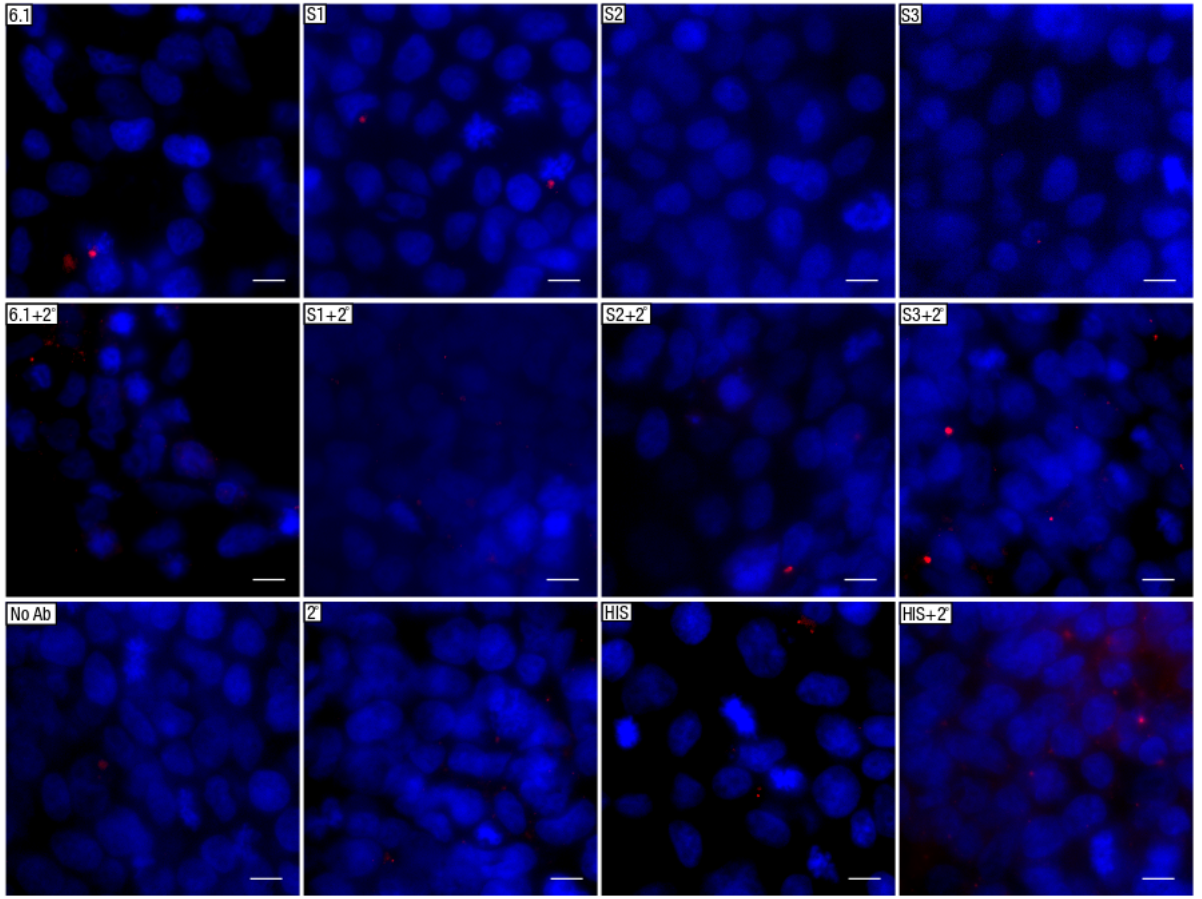

Figure A.2: Immunofluorescence microscopy images of HBx staining in HEK293 cells expressing the HBx negative, pCI-Neo vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in any samples. Magnification is 60x and scale bar = 10μm.

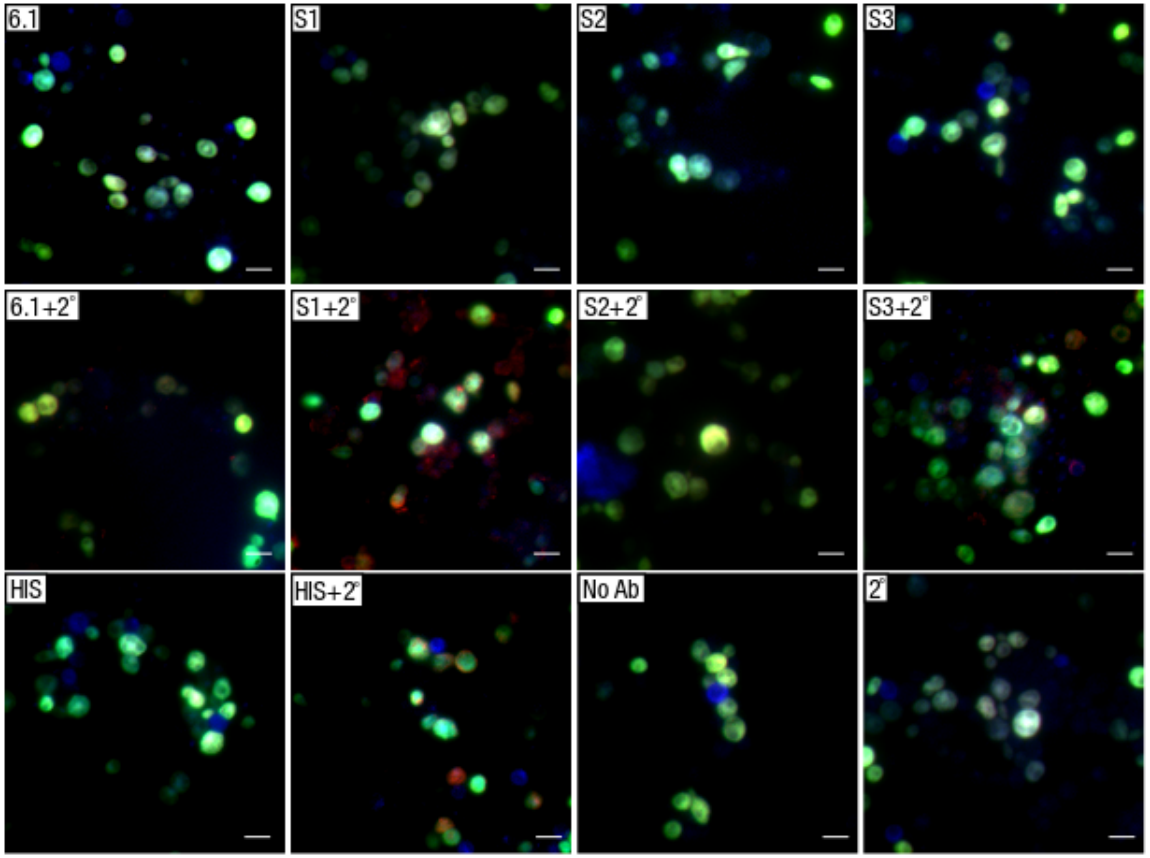

Figure A.3: Immunofluorescence microscopy images of HBx staining in *S. cerevisiae* cells expressing the HBx negative, YEp-GFP vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in samples containing primary or secondary antibody alone. Samples containing both primary and secondary antibodies showed a low presence of Cy3 fluorescence. Magnification is 100x and scale bar = 5μm.

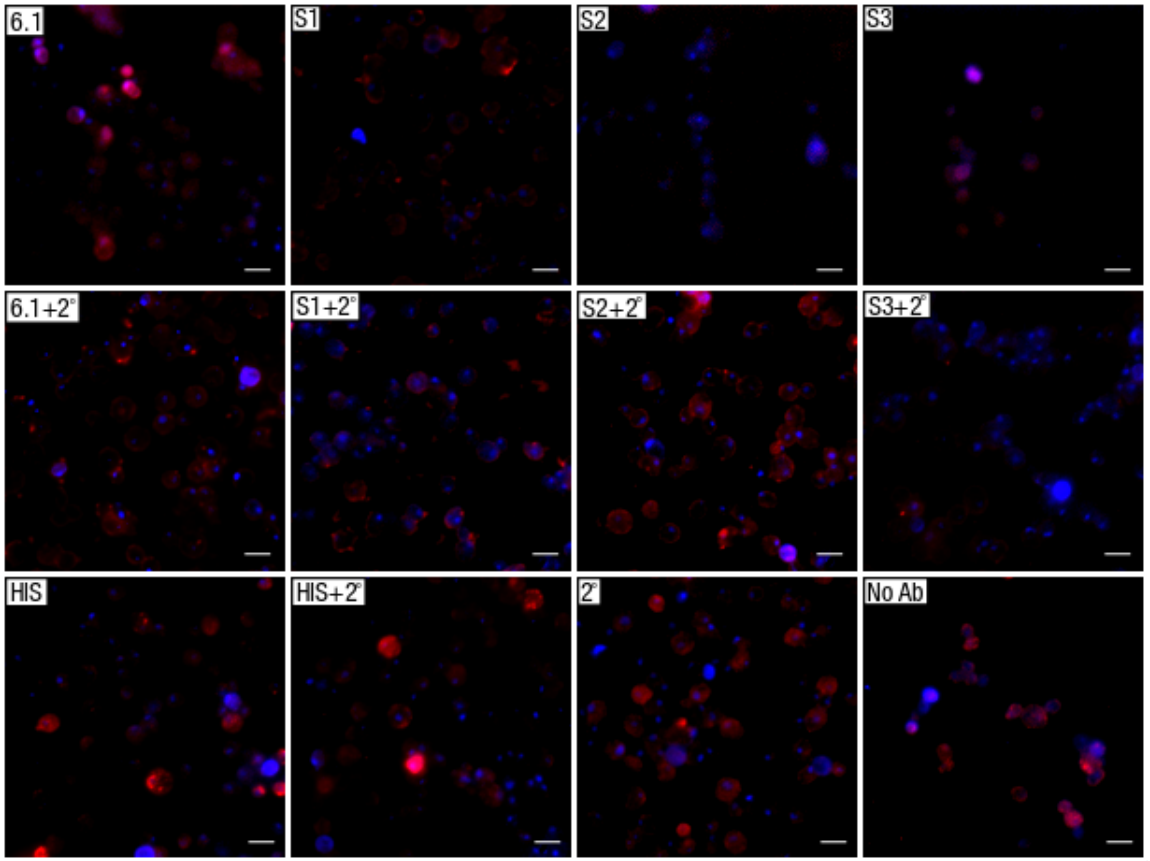

Figure A.4: Immunofluorescence microscopy images of HBx staining in *S. cerevisiae* cells expressing the HBx negative, YEpHF vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in samples containing either primary or secondary antibody only. Samples containing both primary and secondary antibodies displayed a low Cy3 fluorescence. Stronger fluorescence is observed in the anti-HIS plus secondary treatment, as the empty vector is expected to express a short, 6xHIS tagged protein. Magnification is 100x and scale bar = 5μm.

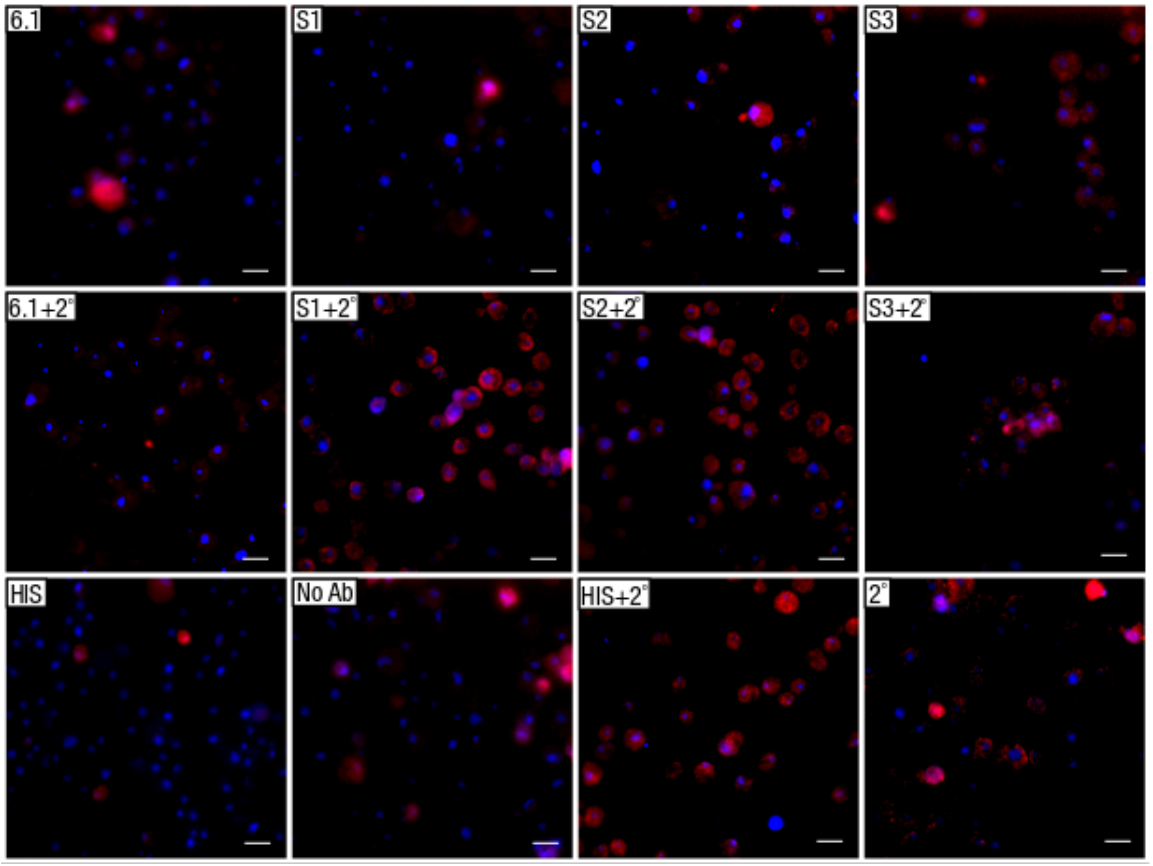

Figure A.5: Immunofluorescence microscopy images of HBx staining in *S. cerevisiae* cells expressing the HBx negative, pYES2 vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. Cy3 fluorescence was not detected in samples only containing primary or secondary antibody. Samples containing both primary and secondary antibodies displayed a low Cy3 fluorescence. Magnification is 100x and scale bar = 5μm.

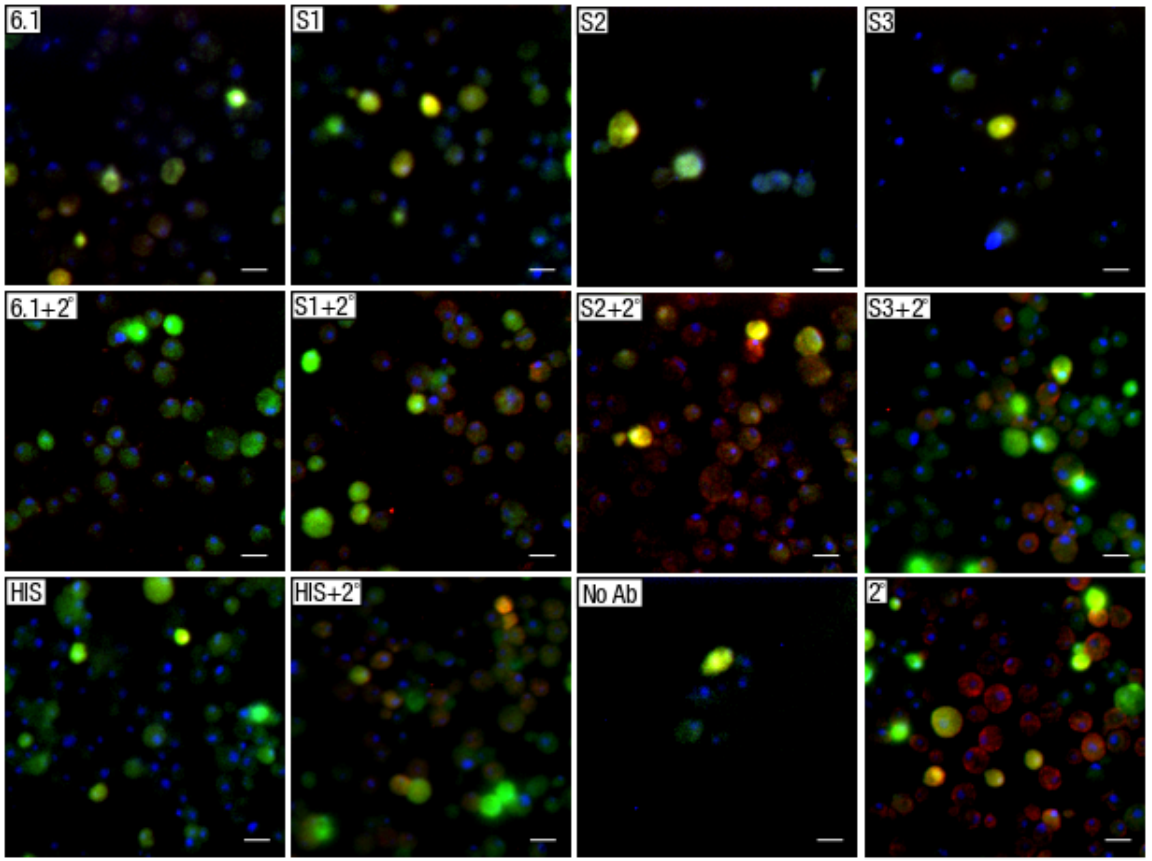

Figure A.6: Immunofluorescence microscopy images of HBx staining in *S. cerevisiae* cells expressing the HBx negative, pYES2-GFP vector. Antibodies specific to each figure are shown in the top left corner. Blue indicates DAPI staining and red indicates anti-HBx antibodies bound to a Cy3 conjugated secondary antibody. A low Cy3 fluorescence was detected in S1,S2, S3 and HIS samples containing both primary and secondary antibody. Magnification is 100x and scale bar = 5μm.

A.1.3 Box and Whisker Key

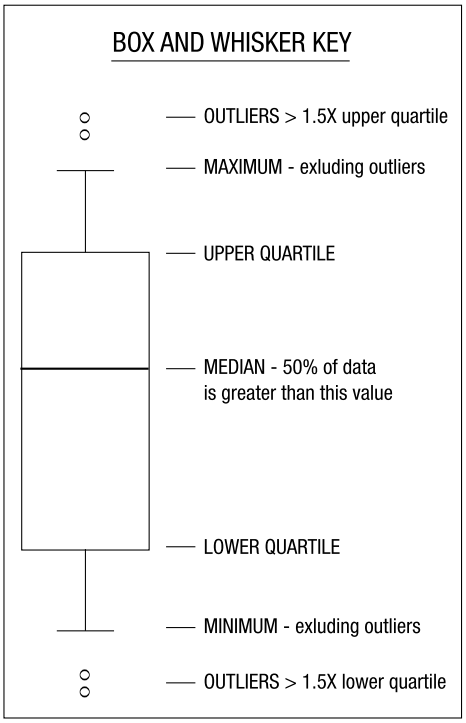

Figure A.7: Key to interpret the data presented in a box and whisker plot. Empty circles outside of the main plot represent outlying data points. The upper bar represents the maximum measured value in the dataset. The upper edge of the box represents the upper quartile, the line going through the middle of the box is the median of the dataset. The bottom half of the plot is a mirror of the top.

A.1.4 Sites Under Selection

Table A.1: Sites in HBx under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. The probability is listed next to each site.

Genotype	Sites Under Selection
genoptye A	5L 0.999, 6Y 0.965, 11S 0.996, 29S 1.000, 30G 1.000, 33L 0.985, 42S 0.898, 83H 0.997, 90P 0.983, 115I 1.000, 118F 0.997, 132S 1.000
genotype B	34L 0.61, 87V 1.000, 90T 0.622, 103T 1.000, 113K 0.675
genotype C	5V 1.000, 86Y 1.000, 107L 1.000, 118I 1.000
genotype D	25C 0.995, 29F 0.998, 32P 0.985, 33L 0.987, 35D 1.000, 36L 0.981, 40S 0.933, 41L 0.987, 42S 0.984, 45T 0.971, 83Q 1.000, 97A 1.000, 115L 0.934
genotype E	29V 1.000, 35D 1.000, 83Q 0.955, 89H 0.962, 117L 0.995, 121T 1.000, 123I 0.988, 138A 0.983
genotype F	12A 0.938, 26R 0.983, 37L 0.991, 40S 0.934, 47D 0.976, 48H 0.999, 97R 0.966, 104E 0.933, 112K 0.967, 121T 0.997, 124K 0.961, 142C 0.977
genotype G	9L 0.584, 12S 0.573, 31P 0.7, 88H 0.689
genotype H	44A 0.815, 87W 0.65

Table A.2: Sites in HBc under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. The probability is listed next to each site.

Genotype	Sites Under Selection
genotype A	96E 1.000
genotype B	39V 0.992, 83L 0.799, 109G 0.621, 119L 0.91, 154Q 0.993, 170R 0.887
genotype C	67S 0.808, 154R 0.946, 179E 0.817
genotype D	27G 0.998, 88D 0.999, 96V 0.972, 102A 0.995
genotype E	5H 0.716, 40T 0.55, 49S 0.509, 88N 0.83, 94A 0.691, 134L 0.796, 148P 0.806
genotype F	11F 0.503, 198S 0.564, 210Q 0.547
genotype G	No Sites
genotype H	2Q 0.517, 42V 0.595, 79N 0.81, 95T 0.527, 98S 0.542, 105E 0.931

Table A.3: Sites in HBs gene under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. The probability is listed next to each site.

Genotype	Sites Under Selection
genoptye A	4W 0.833, 39F 0.899, 44H 0.807, 47Q 0.83, 63V 0.77, 72I 0.751, 76A 0.986
genotype B	49S 0.93, 62T 1.000, 64T 1.000, 95L 0.998, 98Q 0.91, 102L 1.000, 114M 0.960, 115Q 0.997, 134F 1.000, 152S 0.999, 156L 0.993, 157K 1.000, 196N 1.000, 272R 0.982, 275T 0.960, 310Y 1.000, 349F 1.000, 361M 1.000
genotype C	30G 0.988, 48N 1.000, 52V 1.000, 152N 1.000, 154A 0.980, 210I 0.999, 241Q 1.000, 318V 1.000, 329W 0.892
genotype D	27A 0.983, 70Q 0.994, 72A 0.998, 95N 1.000, 134A 0.89, 136H 1.000, 137I 0.998, 146P 0.995, 260T 1.000, 270A 0.905, 276F 0.987
genotype E	13H 1.000, 31R 0.999, 49E 0.992, 76K 0.999, 108M 0.916, 123R 0.994, 125R 1.000, 155P 1.000, 200G 0.977, 201T 0.967, 205L 0.999, 215N 1.000, 266L 0.967, 294S 0.921, 317E 0.991, 353P 0.974
genotype F	135F 0.941, 142E 0.984, 260L 0.906
genotype G	148V 0.997, 149N 0.972, 216V 0.978
genotype H	138F 0.907, 190C 0.754, 201Q 0.767, 333E 0.771, 364I 0.713

Table A.4: Sites in HBp under positive selection as determined using the Bayes Empirical Bayes output of Model 8 from the PAML suite. The probability is listed next to each site.

Genotype	Sites Under Selection
genoptye A	127H 0.986, 170I 0.956, 171K 0.980, 216T 0.997, 247H 0.952, 248S 0.980, 324A 0.995, 554V 0.977
genotype B	15E 0.982, 69K 1.000, 88K 1.000, 98N 0.933, 223Q 0.946, 312L 0.947, 323E 0.938, 331C 0.975, 657N 0.992, 661N 0.973, 781L 0.998, 813R 0.932
genotype C	206C 0.985, 228K 0.985, 260T 0.954, 285A 0.991, 288L 0.998, 381H 0.999, 525M 0.938, 541S 0.986, 556N 0.994, 584Q 1.000, 587H 0.957, 708K 0.997, 767R 1.000, 791D 0.999, 804R 1.000
genotype D	10R 0.932, 114K 0.998, 201C 0.995, 229V 0.986, 233A 0.916, 257L 0.977, 258Y 0.979, 269S 0.927, 275S 1.000, 285F 0.983, 447F 0.909, 470L 0.934, 538P 0.997, 587R 1.000, 592E 0.999, 683A 0.959, 694S 1.000, 719L 0.994, 772S 1.000
genotype E	162C 1.000, 212P 0.987, 221R 0.993, 231R 0.990, 233G 0.919, 249K 0.982, 282T 0.965, 541S 0.981, 572W 0.968, 580K 0.972, 581L 1.000, 644L 0.968, 677M 0.995, 756I 0.988, 757Y 0.995, 796R 0.970
genotype F	30G 0.914, 121L 0.967, 198T 0.901, 261S 0.955, 301P 0.908, 314F 0.995, 486L 0.915, 579N 0.993, 646T 0.972, 729N 0.924
genotype G	305P 0.895, 329S 0.960, 330R 0.982, 769R 0.981
genotype H	35H 0.918, 100K 0.949, 118V 0.926,3 16F 0.917, 591I 0.984, 671L 0.93, 786Y 0.936

Appendix B

Standard Laboratory Methods

B.1 Bacterial growth, storage and transformations

B.1.1 Bacterial growth and storage

The host bacterial strain used in normal cloning work was *Escherichia coli* DH5$\alpha^{(269)}$, whereas protein expression used the BLR-DE3 strain. Cultures were grown with shaking (350 rpm) in Erlenmeyer flasks at 37 °C in a New Brunswick Series 25 orbital incubator (New Brunswick Scientific Co, Inc, NJ, USA) in Luria Bertani (LB) broth (B.3.2) or on LB agar (LA) plates supplemented with 200 μg/mL ampicillin antibiotic (Sigma-Aldrich, MO , USA). *E.coli* DH5α transformants were stored in 33 % glycerol (Merck, Darmstadt, Germany) at -80 °C.

B.1.2 Preparation of chemically competent *E.coli*

E. coli cells were grown overnight in 2 mL of LB, and diluted 1:100 in 100 mL of LB (see B.3.2) until log phase was reached, as measured by an OD_{600} of 0.4. Cells were chilled for 20 minutes on ice, and were pelleted by centrifugation for ten minutes at 4 °C and 59.5 x g in a refrigerated centrifuge (Eppendorf 5415D/R, Eppendorf, Hamburg, Germany). The pellet was resuspended in 10 mL of ice cold, sterile transformation buffer (60 m*M* $CaCl_2$, 15 % glycerol, 10 m*M* PIPES pH 7.0 at temperature; Merck, Darmstadt, Germany) and incubated for 20 minutes on ice. The suspension was then pelleted by centrifugation for 10 minutes at 4 °C and 59.5 x g. The pellets was resuspended in 2 mL of ice cold transformation buffer and 200 μL aliquots were stored at -80 °C.

B.1.3 Bacterial transformations

Plasmid DNA (0.5 μg/μL) was added to a 100 μL aliquot of transformation competent cells (thawed on ice for 5 minutes). This was incubated on ice for 30 minutes before being heat shocked for 90 seconds at 42 °C. Following the addition of 1 mL of LB, cells were incubated for 60 minutes at 37 °C

with shaking. This allowed time for the phenotypic expression of antibiotic resistance genes, prior to plating onto ampicillin LA plates.

B.1.4 Transformation and growth of *S. cerevisiae*

Transformation of plasmids into *S. cerevisiae* was done using an adaptation of the method described in Chen *et al*[270]. Briefly, 800 μL of stationary phase yeast was centrifuged at 16 000 x g, after which the supernatant was removed. The pellet was resuspended in a mixture that had been preheated to 45 °C, comprising 50 ng - 1 μg of plasmid DNA, 5 μL of 10 μg/mL salmon sperm ssDNA that had been pre-sheared, boiled and chilled. To this, 100 μL of one step buffer [200 m*M* lithium acetate, 40 % PEG 4000 and 100 m*M* dithiothreitol (DTT)] was added, and the pellet was vortexed to resuspend the cells. The transformation mixture was incubated at 45 °C for 30 minutes. Cells were pelleted by centrifugation at 3 500 x g for 3 minutes at room temperature and resuspended in 200 μL of YPD, and plated on appropriate SC agar plates.

B.2 Plasmid DNA preparation, electrophoresis and extraction

B.2.1 Plasmid DNA preparation

Plasmid DNA was prepared using three commercially available kits (Qiagen, Hilden, Germany) that differed according to the concentration of DNA required mini-preps (150 ng/μL plasmid DNA), midi-preps (800 ng/μL of plasmid DNA) and maxi-preps (5 μ/gμL of plasmid DNA). Transformed *E.coli* DH5α or BLR-DE3 cultures (two mL for mini-preps, 100 mL for midi-preps and 250 mL for maxi-preps) were grown in LB overnight. Cells were pelleted by centrifugation for 5 minutes at 4000 x g (midi and maxi preps) or 16 000 x g for one minute (mini preps). Pellets were resuspended in 300 μL (mini prep), 10 mL (midi prep) or 15 mL (maxi prep) of Buffer P1 (50 m*M* Tris-HCl pH 8.0, 10 m*M* EDTA and 100 μg/ml RNAse A). An equal volume of Buffer P2 [200 m*M* NaOH, 1 % SDS (w/v)] was added and mixed by inversion for one minute. An equal volume of pre-chilled Buffer P3 (3.0 M potassium acetate pH 5.5) was added (for midi and maxi preps; 300 μL for mini preps) and mixed by inversion for one minute prior to centrifugation for 20 minutes at 4000 x g for midi and maxi preps or 16 000 x g for mini preps.

For mini preps, the supernatant was pipetted into a new 2 mL tube, and an equal volume (900 μL) of isopropanol was added. The samples were vortexed for 5 seconds, and chilled for 20 minutes at -20 °C, after which they were centrifuged for 10 minutes. The supernatant was discarded, and the pellet was washed in 250 μL of 70 % ethanol, and centrifuged again for 5 minutes. The supernatant was discarded, and the pellets left to air dry for 10 minutes.

For midi and maxi preps, the supernatant was poured into a QIAtip binding column that had

been equilibrated with 4 mL (midi prep) or 10 mL (maxi prep) of Buffer QBT [750 m*M* NaCl, 50 m*M* MOPS pH 7.0, 15 % isopropanol (v/v), 0.15 % Triton® X-100 (v/v)] using gravity flow. The column was washed with 10 mL (midi prep) or 30 mL (maxi prep) of Buffer QC [1.0 M NaCl, 50 m*M* MOPS pH 7.0, 15 % isopropanol (v/v)]. Plasmid DNA was eluted into a new 50 mL tube using 9 mL (midi prep) or 15 mL (maxi prep) of Buffer QF [1.25 M NaCl, 50 m*M* Tris-HCl pH 8.5, 15 % isopropanol (v/v)] followed by the addition of 6.3 mL (midi prep) or 10.5 mL (maxi prep) of isopropanol (Merck, Darmstadt, Germany). Samples were chilled for 2 hours at -20 °C and centrifuged for 30 minutes at 4 °C and 8000 x g. Pellets were air-dried and resuspended in 250 μL or 500 μL of Buffer TE (10 m*M* Tris-HCl pH 8.0, 1 m*M* EDTA pH 8.0). Plasmid DNA was subjected to spectrophotometry using a NanoDrop Spectrophotometer (Thermo Fisher Scientific, Inc, MA, USA) to determine the concentration (at a wavelength of 260 nm, 1 absorbance unit at A260 = 50 μg/mL dsDNA and purity ratio A260/A280 $>$1.8).

B.2.2 Phenol/chloroform extraction of DNA

The phenol:chloroform:isoamyl alcohol method of nucleic extraction was used to recover DNA for ligation or sequencing analysis following enzymatic digestion or PCR. The recipe for SS-phenol used in this method was taken from[271]. DNA solutions were brought to a volume of 100 μL with nuclease-free TE buffer (Qiagen, Hilden, Germany). A 100 μL volume of 25:24:1 phenol:chloroform:isoamyl alcohol mixture added to the samples which were briefly mixed to form an emulsion. The emulsion was centrifuged for 2 minutes at 12 000 x g. Approximately 100 μL of the aqueous phase was transferred to a new microfuge tube followed by the addition of one volume of chloroform. The samples were briefly mixed and centrifuged for 2 minutes at 12000 x g. The aqueous phase was removed and 0.1 volumes of 3 M sodium acetate pH 5.2 and 2.5 volumes of 100 % ethanol were added. Samples were chilled at -70 °C for 15 minutes followed by centrifugation for 15 minutes at 13 000 x g and 4 °C. The pellet was washed with 70 % ethanol, re-centrifuged for 10 minutes, after which samples were air-dried for 10 minutes at 37 °C to remove residual ethanol. Pellets were resuspended in an appropriate volume of nuclease-free water. All DNA extracted from PCRs or agarose gels was quantified using comparative densitometry with 0.5 μg of DNA ladder (O'GeneRuler TM Mix, Fermentas, WI, USA).

B.2.3 Agarose gel electrophoresis

To make a 0.8 % agarose gel, 0.8 g of agarose powder (Sigma-Aldrich, MO, USA) was dissolved in 100 mL of 1 x TAE buffer (Tris base 40 m*M*, 20 m*M* Na-acetate, 1 m*M* EDTA) and heated until the agarose had dissolved. Prior to pouring, one μL per 50 mL gel of ethidium bromide (10mg/mL - Sigma, Germany) was added. DNA samples were mixed with 0.2 volume of loading dye (1.67

m*M* Tris-HCl pH 7.6, 0.005 % bromophenol blue, 0.005 % xylene cyanol FF, 10 % glycerol, 10 m*M* EDTA, Fermentas, WI, USA) and electrophoresed parallel to 5 μL of a DNA molecular weight marker (O'GeneRuler TM Mix, Fermentas, WI, USA) that had been loaded into the wells of a set agarose gel. Gels were subjected to electrophoresis at 100 volts.

B.2.4 DNA purification from agarose gels

DNA fragments were purified from agarose gels by placing the excised gel fragment inside a 0.5 mL PCR tube that had been punctured at the bottom using a hot 22G needle. Nylon fishtank wool was placed inside the 0.5 mL PCR tube to act as a sieve. The gel fragment was placed inside the 0.5 mL PCR tube, which was placed inside a 2.0 mL microfuge tube and centrifuged at 12 000 x g for 2 minutes. DNA purification proceeded as described under phenol/chloroform extraction.

B.2.5 RNA extraction from adherent mammalian cells seeded in 10 cm dishes

Forty eight hours post-transfection mammalian HEK293 or Huh7 cells that had been cultured in 10 cm dishes (NuncTM, AEC-Amersham (Pty) Ltd, South Africa) were washed in PBS (Gibco, UK), resuspended in one mL of TriReagent® (Sigma, Germany) and transferred to a microfuge tube with 300 μL of chloroform (Merck, Germany) followed by ten seconds of vortexing. Samples were centrifuged for 30 minutes at 16 000 x g and 4 °C in a refrigerated centrifuge (Eppendorf 5415D/R; Eppendorf, Germany). The upper aqueous phase was carefully transferred to a new microfuge tube, an equal volume ($\sim$ 300 μL) of isopropanol (Merck, Germany) was added and samples were centrifuged for 30 minutes at 16 000 x g and 4 °C. The supernatant was discarded and the pellets air dried before being resuspended in 50 μL formamide (Sigma-Aldrich, MO , USA). RNA was subjected to spectrophotometry using a NanoDrop Spectrophotometer to determine the concentration (at a wavelength of 260 nm, 1 absorbance unit (A260) = 40 μg/mL RNA) and purity (A260/A280 $>$2).

B.2.6 *S. cerevisiae* RNA extraction

RNA was extracted from yeast cultures by resuspending centrifuged pellets in 400 μL of 1x TES buffer (10 m*M*, Tris-HCI pH 7.5 10 m*M* EDTA, 0.5% SDS) and adding an equal volume of acid phenol. The cells were vortexed for 10 seconds, and incubated for 1 hour at 65 °C, with occasional vortexing. The samples were centrifuged at maximum speed in a 4 °C microfuge for 5 minutes. The upper phase was transferred into a new tube, and the sample re-extracted with an equal volume of acid phenol. The sample was treated in the same way as mentioned in B.2.2. Samples were resuspended in formamide (Sigma-Aldrich, MO , USA) and stored at -20 °C.

B.2.7 DNA Sequencing

Plasmids or PCR products were sequenced using the BigDye® Terminator Kit 3.1 (Applied Biosystems, CA, USA). The protocol of the manufacturer was followed, except that a half volume reaction was used. The cleaning reaction was performed as per manufacturers instructions, except that the volume of the sequencing reaction was made up to 100 μL with dH_2O before cleaning.

B.3 Cloning Protocols

Cloning conditions for the myriad of constructs made were standardised, as they were found to work in almost every case. Backbone DNA was prepared from bulk preparations (B.2) and 20 μg of backbone plasmid DNA was used in each digestion. 2 units of appropriate restriction enzyme per reaction were used as restriction digests were done overnight at 37 °C.

If restriction digests used 2 different enzymes and the cleavage sites showed ligation compatibility or the restriction digest used only 1 restriction enzyme, the DNA backbone would be dephosphorylated using 1 unit FastAP (Fermentas, WI, USA) per μg of DNA for 10 minutes at 37 °C. If the cleavage sites showed no ligation potential, or blunt end cloning was involved, the backbone DNA was not dephosphorylated to increase the likelihood of cloning of desired fragments. All DNA was cleaned using the phenol:chloroform method (B.2.2) after the correct DNA bands had been excised from an agarose gel (B.2.4. Following comparative densitometry to determine the concentration of purified DNA, a ligation reaction would bet set up as described in the main body of text.

B.3.1 pTZ57R/T TA Cloning Protocol

PCR products for cloning into a TA vector, were inserted into pTZ57R/T of the InsTAclone™ kit (Fermentas, WI, USA). Briefly, a 20 μL ligation reaction comprised the following - 1 μL pTZ57R/T (55ng), 1 μL T4 DNA Ligase (5 units), 2 μL 10X Ligation Buffer, 1 μL PCR product, and 15 μL nuclease free water. The ligation mixture was incubated at 6 °C overnight, after which 10 μL was transformed into competent *E. coli* as described B.1.3.

B.3.2 Preparations for Bacterial Cultures

B.3.2.1 ZYM-5052 Media

1 % g Tryptone (Oxoid, U.K.), 0.5 % Yeast Extract (Oxoid), 25 m*M* Na_2HPO_4, 25 m*M* KH_2PO_4, 50 m*M* NH_4Cl, 5m*M* Na_2SO_4, 2 m*M* $MgSO_4$, 1m*M* $FeCl_3$, 0.5 % glycerol and 0.05 % glucose. For bacterial protein induction, 0.5 % Lactose was added, this media is ZYM-505. Media was made up to a total volume of 1 litre with pure H_2O. Medium was sterilized by autoclaving for 20 minutes on a liquid cycle and stored at room temperature.

B.3.2.2 Luria-Bertani Broth (LB)

B.3.2.3 Agar Media

Bacterial plates contained 1.5 % (w/v) Agar (Oxoid) in LB medium, whereas yeast plates contained 2.0 % (w/v) agar.

B.3.2.4 Antibiotic Stock Solutions

100 mg/ml of ampicillin (Roche, Germany) in 50 % ethanol; 35 mg/ml of chloramphenicol in isopropanol. Both antibiotic solutions were stored at $-20\,^{\circ}$C.

B.3.2.5 IPTG and X-galactosidase

For induction of T7 protein expression and blue-white colony screening, IPTG (Roche, Germany) solutions were stored at $-20\,^{\circ}$C. 200 mg of X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) was dissolved in 10 mL of N,N-dimethylformamide and stored at $-20\,^{\circ}$C in foil. 40 μL of X-gal and 4 μL of IPTG (200 mg/ml aqueous solution, Roche, Basel, Switzerland) were spread on Ampicillin positive LB agar plates and left to air dry for 15 minutes at $37\,^{\circ}$C prior to plating.

B.4 Preparations for Eukaryotic Cell Cultures

B.4.1 Maintenance and Transfection of plasmid DNA into mammalian cells

The human embryonic kidney (HEK293) cell line was maintained using Dulbecco's Modified Eagle's Medium (DMEM, Lorenza Biosciences, Basel, Switzerland) supplemented with heat-inactivated 10% fetal calf serum (FCS, Invitrogen, CA, USA) at $37\,^{\circ}$C and 5% CO_2 in a Forma Series II 3110 Water Jacketed CO_2 incubator (Thermo Fisher Scientific, Inc, MA, USA).

Transfection of plasmid DNA into HEK293 cells was performed by either the Calcium-Phosphate method for 10cm tissue culture dishes, or Lipofectamine2000 (Invitrogen, CA, USA), if 12-48 well tissue culture plates were used. The Lipofectamine2000 to plasmid DNA ratio was maintained at 1:1, and the maximum quantity of DNA recommended was used for each transfection. There were no other modifications to this protocol.

B.4.1.1 Antibiotic Solutions for tissue culture

Penicillin in pure H_2O (100 000 U/ml, Invitrogen-Gibco-BRL, CA USA) Streptomycin in pure H_2O (100 mg/ml, Invitrogen-Gibco-BRL, CA USA) The antibiotic solutions were filter-sterilized using 0.2 μm filters and stored in 500μL aliquots at $-20\,^{\circ}$C.

B.5 General Solutions

B.5.1 DNA/RNA Electrophoresis Buffers

B.5.1.1 (50x) Tris-acetate- EDTA (TAE):

2 M Tris-acetate (pH 8.0), 50 m*M* EDTA. To prepare 500 ml of a 50 x stock, 121g Tris base, 28.5 ml glacial acetic acid and 50 ml 0.5 M EDTA (pH 8.0) were mixed in 400 ml of pure H_2O. The solution was made up to 500 ml with pure H_20 and stored at room temperature.

B.5.1.2 (10x) Tris-borate-EDTA (TBE):

0.9 M Tris-borate (pH 8.0), 200 m*M* EDTA A 10 x stock of TBE (1 litre) was prepared by mixing 54 g, 27.5 g Boric acid and 20 ml 0.5 M EDTA (pH 8.0) in 900 ml of pure H_2O. The solution was made up to 1 litre with pure H_2O and stored at room temperature.

B.5.1.3 DNA loading Buffer

6 x DNA loading buffer contained 3 ml of glycerol and 25 mg of bromophenol blue (Merck, Germany) made up to 10 ml with pure H_2O. DNA loading buffer was stored at room temperature.

B.5.1.4 RNA Loading Buffer

A 2 x RNA loading buffer was made up by mixing 990 μL of deionized formamide (Saarchem, South Africa), 2 μL of 500 m*M* EDTA (pH 8.0), 2.5 mg of bromophenol blue and 2.5 mg of Xylene Cyanol FF (Merck, Germany). RNA loading buffer was stored at $-20\,^{\circ}$C in 100 μL aliquots.

B.5.2 SDS-PAGE and Western blotting

SDS-PAGE and Western blotting protocols were performed as described in Unit 10.8.1 in Current Protocols in Molecular Biology[272].

Immunoblot Buffers

1. (5x) SDS-PAGE Tank Buffer: (125 m*M* Tris (pH 8.3), 0.958 M Glycine, 0.5 % SDS).
2. A 10 x stock of tank buffer was prepared by dissolving 30 g Tris base, 144 g Glycine (BDH, Germany) and 10 g SDS in 1.5 litres of pure H_2O. The solution was made up to 2 litres with pure H_2O and stored at room temperature.
3. A-3.5 SDS-PAGE Loading Buffer: 125 m*M* Tris (pH 6.8), 4 % SDS, 20 % glycerol, 200 m*M* DTT, 0.02 % Bromophenol Blue).

4. A 2 x Loading buffer was prepared by mixing the following components in a total volume of 10 ml of pure H_2O: 0.15 g Tris base, 4.0 ml of 10 % SDS, 2.0 ml Glycerol, 2.0 mg Bromophenol Blue (Merck, Germany) and 0.31 g Dithiothreitol (DTT, Roche, Germany). 500 μL aliquots were stored at $-20\,^{\circ}$C.

5. Transfer Buffer: 125 m*M* Tris (pH 8.3), 0.958 M glycine, 0.5 % SDS, 20 % methanol.

6. Tris-Buffered Saline (TBS): 25 m*M* Tris (pH 7.4), 150 m*M* NaCl, TBS (1 litre) was prepared by dissolving 3.0 g Tris base, 0.2 g KCl and 9.0 g NaCl in 900 ml of pure H_2O. The pH of the solution was adjusted to 7.4 with HCl and made up to 1 liter with pure H_2O. TBS-T consisted of TBS with 0.1 % Tween-20 (Saarchem, South Africa).

7. Blocking buffer consisted of T-TBS with 5 % fat-free milk powder.

B.5.2.1 Staining solution

1. Coomassie Brilliant Blue: 0.025% Coomassie Blue, 40 % methanol, 7 % acetic acid.

2. Staining solution contained 0.5 g Coomassie Brilliant Blue R250 (Merck, Germany), 800 ml methanol and 140 ml glacial acetic acid made up to a total volume of 2 litres with pure H_2O, and was stored at room temperature.

3. Destaining Solution contained 10 % methanol and 10 % acetic acid.

B.6 General ELISA Protocol

1. The antigen is diluted to a final concentration between 5 and 10 μg/ml in Coating Buffer (alt. PBS). Coating buffer comprises (50 *m* M carbonate/bicarbonate buffer, pH 9.6) made up as follows: 15 m*M* Na_2CO_3, 35 m*M* $NaHCO_3$, 0.05 % azide, 0.05 % Tween 20.

2. Dispense 100 μl into each Elisa plate well. Incubate at 37 $^{\circ}$C for 2 hrs or in the fridge overnight.

3. Discard coating solution. Wash 3x with washing solution (0.9% NaCl + 0.05% Tween 20; alternate. PBS). Dry plate by inversion and tapping onto towel paper.

4. Block possible remaining binding sites in wells with 150μl per well of Dilution Buffer : 50m*M* Tris, pH 7.4 (alternate PBS) 2% BSA or non-fat milk powder, 0.05% Tween 20, 0.02% Na-azide. Incubate at room temperature 30 – 40 minutes.

5. Discard solution and dry as before. Add 100μl clonal supernatant and 50μl Dilution Buffer to each corresponding well. Incubate for 30 minutes at room temperature.

6. Discard solution; wash 3x with Washing Solution; dry as before. Add 100μl peroxidase conjugated anti-mouse Ig at the appropriate dilution in PBS. Incubate at room temperature for 30 minutes.

7. Discard solution; wash 3x with Washing Solution and 1X with 100m*M* Citrate Buffer, pH 5.0 made up as follows: citric acid.H_2O 46m*M*, Na_2HPO_4 67m*M*. Discard and dry as before.

8. Add 10μl freshly prepared Substrate Solution : 20mg OPD.2HCl in 30ml 100m*M* citrate buffer + 20μl 30% H_2O_2. Check for colour development. The reaction may be stopped by the addition of 150μl 1M H_2SO_4. Colour intensity was read at 496nm.

Appendix C

Python Computer Code

An example of the Python code used to extract genotype specific gene sequences from a database of re-genotyped and re-annotated HBV sequences. Additional scripts were written for a variety of purposes but have not been shown for the sake of brevity.

```
## Import the Sequence function of BioPython
from Bio import SeqIO
subtype_list = (([ ],[ ],[ ],[ ]),([ ],[ ],[ ],[ ]),
        ([ ],[ ],[ ],[ ]),([ ],[ ],[ ],[ ]),
        ([ ],[ ],[ ],[ ]),([ ],[ ],[ ],[ ]),
        ([ ],[ ],[ ],[ ]),([ ],[ ],[ ],[ ]))

## subtype_list is a list within a list and is meant to hold the various records
## from the the input file below which contains a mixture of genes and genotypes.
## The order of the list is based on the orf_dic and subtype_dic directly below.
## i.e. The first position will be subtype A records, and the first entry is the
## polymerase, followed by X, surface and precore-core etc.

orf_dic = (['polymerase'],['x'],['surface'],['precore-core'])
subtype_dic = (['A'],['B'],['C'],['D'],['E'],['F'],['G'],['H'])
sequence = ""

from Bio.SeqRecord import SeqRecord

## optional function to translate the DNA sequence into protein
def make_protein_record(nuc_record):
    """Returns a new SeqRecord with the translated sequence (default table)."""
    return SeqRecord(seq = nuc_record.seq.translate(to_stop=False, stop_symbol="*"), \
                     id = "trans_" + nuc_record.id, \
                     description = "translation of CDS, using default table")

## function to write records to file - using genotype and gene as a filename
def write_records(receive):
    for i,v in enumerate(subtype_dic):
        for x,y in enumerate(orf_dic):
```

```
            filename = "genotye_" + str(v[0]) + "_" + str(y[0]) + ".fasta"
            handle = open(filename, 'w')
            for entries in receive[i][x]:
                output = str(entries)
                handle.write(output)
            handle.close()

            proteins = (make_protein_record(nuc_rec) for nuc_rec in \
            SeqIO.parse(filename, "fasta"))
            filename = "genotye_" + str(v[0]) + "_" + str(y[0]) + "_translations.fasta"
            SeqIO.write(proteins, filename, "fasta")

for records in SeqIO.parse("cts.gb", "gb"):
    print records.name
    #gives position (i) and value (v) of each entry in subtype_dic
    for i,v in enumerate(subtype_dic):

        if records.features[0].qualifiers['genotype'] == v:
            for orfs in records.features:
                if orfs.type == "CDS":
                    for x,y in enumerate(orf_dic):
                    #gives position (x) and value (y) of each entry in orf_dic
                        if orfs.qualifiers['product']  == y:
                            if len(orfs.sub_features) == 2:
                                for sub_f in orfs.sub_features:
                                    sequence = sequence + \
                    records.seq[sub_f.location.nofuzzy_start:sub_f.location.nofuzzy_end]
                            if len(orfs.sub_features) == 0:
                                sequence = records.seq[orfs.location.nofuzzy_start:orfs.location.↩
                                    nofuzzy_end]
                            sequence = ">" + str(records.annotations['accessions'][0]) + "\n" + ↩
                                sequence + "\n"
                            subtype_list[i][x].append(sequence)
                            sequence = ""
write_records(subtype_list)
```

Bibliography

[1] F. O. MacCallum, "Homologous serum hepatitis," *The Lancet*, vol. 250, no. 6480, pp. 691 – 692, 1947. Originally published as Volume 2, Issue 6480.

[2] B. S. Blumberg, H. J. Alter, and S. Visnich, "A "new" antigen in leukemia sera.," *JAMA*, vol. 191, pp. 541–546, Feb 1965.

[3] K. Okochi and S. Murakami, "Observations on Australia antigen in Japanese," *Vox Sang*, vol. 15, no. 5, pp. 374–385, 1968.

[4] D. S. Dane, C. H. Cameron, and M. Briggs, "Virus-like particles in serum of patients with Australia-antigen-associated hepatitis.," *Lancet*, vol. 1, pp. 695–698, Apr 1970.

[5] W. S. Robinson, "Molecular events in the pathogenesis of hepadnavirus-associated hepatocellular carcinoma.," *Annu Rev Med*, vol. 45, pp. 297–323, 1994.

[6] F. Trevisani, P. E. D'Intino, A. M. Morselli-Labate, G. Mazzella, E. Accogli, P. Caraceni, M. Domenicali, S. D. Notariis, E. Roda, and M. Bernardi, "Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of hbsag and anti-hcv status.," *J Hepatol*, vol. 34, pp. 570–575, Apr 2001.

[7] World Health Organisation, "Hepatitis B," 2008.

[8] S. T. Goldstein, F. Zhou, S. C. Hadler, B. P. Bell, E. E. Mast, and H. S. Margolis, "A mathematical model to estimate global hepatitis B disease burden and vaccination impact.," *Int J Epidemiol*, vol. 34, pp. 1329–1339, Dec 2005.

[9] C. E. Stevens, R. P. Beasley, J. Tsui, and W. C. Lee, "Vertical transmission of hepatitis B antigen in Taiwan.," *N Engl J Med*, vol. 292, pp. 771–774, Apr 1975.

[10] D. Daniels, S. Grytdal, A. Wasley, C. for Disease Control, and P. (CDC), "Surveillance for acute viral hepatitis - United States, 2007.," *MMWR Surveill Summ*, vol. 58, pp. 1–27, May 2009.

[11] N. Moolla, M. Kew, and P. Arbuthnot, "Regulatory elements of hepatitis B virus transcription.," *J Viral Hepat*, vol. 9, pp. 323–331, Sep 2002.

[12] C. Seeger and W. S. Mason, "Hepatitis B virus biology.," *Microbiol Mol Biol Rev*, vol. 64, pp. 51–68, Mar 2000.

[13] X. Lu and T. Block, "Study of the early steps of the hepatitis B virus life cycle.," *Int J Med Sci*, vol. 1, no. 1, pp. 21–33, 2004.

[14] A. S. F. Lok and B. J. McMahon, "Chronic hepatitis B," *Hepatology*, vol. 45, no. 2, pp. 507–539, 2007.

[15] A. McLachlan, *Molecular biology of the hepatitis B virus*. CRC Press, 1991.

[16] K. Ando, L. Guidotti, S. Wirth, T. Ishikawa, G. Missale, T. Moriyama, R. Schreiber, H. Schlicht, S. Huang, and F. Chisari, "Class i-restricted cytotoxic T lymphocytes are directly cytopathic for their target cells *in vivo*," *The Journal of Immunology*, vol. 152, no. 7, pp. 3245–3253, 1994.

[17] L. G. Guidotti, T. Ishikawa, M. V. Hobbs, B. Matzke, R. Schreiber, and F. V. Chisari, "Intracellular inactivation of the hepatitis B virus by cytotoxic T lymphocytes.," *Immunity*, vol. 4, pp. 25–36, Jan 1996.

[18] L. G. Guidotti, R. Rochford, J. Chung, M. Shapiro, R. Purcell, and F. V. Chisari, "Viral clearance without destruction of infected cells during acute hbv infection.," *Science*, vol. 284, pp. 825–829, Apr 1999.

[19] D. Ganem and A. M. Prince, "Hepatitis b virus infection–natural history and clinical consequences.," *N Engl J Med*, vol. 350, pp. 1118–1129, Mar 2004.

[20] R. P. Beasley, L. Y. Hwang, C. C. Lin, and C. S. Chien, "Hepatocellular carcinoma and hepatitis B virus. a

prospective study of 22 707 men in Taiwan.," *Lancet*, vol. 2, pp. 1129–1133, Nov 1981.

[21] K. Sakuma, N. Saitoh, M. Kasai, H. Jitsukawa, I. Yoshino, M. Yamaguchi, K. Nobutomo, M. Yamumi, F. Tsuda, and T. Komazawa, "Relative risks of death due to liver disease among Japanese male adults having various statuses for hepatitis B s and e antigen/antibody in serum: a prospective study.," *Hepatology*, vol. 8, no. 6, pp. 1642–1646, 1988.

[22] C. Brechot, C. Pourcel, A. Louise, B. Rain, and P. Tiollais, "Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma.," *Nature*, vol. 286, pp. 533–535, Jul 1980.

[23] P. R. Chakraborty, N. Ruiz-Opazo, D. Shouval, and D. A. Shafritz, "Identification of integrated hepatitis B virus DNA and expression of viral RNA in an HBsAg-producing human hepatocellular carcinoma cell line.," *Nature*, vol. 286, pp. 531–533, Jul 1980.

[24] G. Fourel, C. Trepo, L. Bougueleret, B. Henglein, A. Ponzetto, P. Tiollais, and M. A. Buendia, "Frequent activation of N-myc genes by hepadnavirus insertion in woodchuck liver tumours.," *Nature*, vol. 347, pp. 294–298, Sep 1990.

[25] T. Hsu, T. M^{f}or^{f}oy, J. Etiemble, A. Louise, C. Trépo, P. Tiollais, and M. A. Buendia, "Activation of c-myc by woodchuck hepatitis virus insertion in hepatocellular carcinoma.," *Cell*, vol. 55, pp. 627–635, Nov 1988.

[26] A. Dejean, L. Bougueleret, K. H. Grzeschik, and P. Tiollais, "Hepatitis B virus DNA integration in a sequence homologous to v-erb-a and steroid receptor genes in a hepatocellular carcinoma.," *Nature*, vol. 322, no. 6074, pp. 70–72, 1986.

[27] I. C. Hsu, R. A. Metcalf, T. Sun, J. A. Welsh, N. J. Wang, and C. C. Harris, "Mutational hotspot in the p53 gene in human hepatocellular carcinomas.," *Nature*, vol. 350, pp. 427–428, Apr 1991.

[28] B. Bressac, M. Kew, J. Wands, and M. Ozturk, "Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa.," *Nature*, vol. 350, pp. 429–431, Apr 1991.

[29] Y. B. Wang, L. Z. Lan, B. F. Ye, Y. C. Xu, Y. Y. Liu, and W. G. Li, "Relation between geographical distribution of liver cancer and climate-aflatoxin B1 in China.," *Sci Sin B*, vol. 26, pp. 1166–1175, Nov 1983.

[30] P. Merle and C. Trepo, "Molecular mechanisms underlying hepatocellular carcinoma.," *Viruses*, vol. 1, pp. 852–872, Dec 2009.

[31] K. H. Buetow, V. C. Sheffield, M. Zhu, T. Zhou, F. M. Shen, O. Hino, M. Smith, B. J. McMahon, A. P. Lanier, and W. T. London, "Low frequency of p53 mutations observed in a diverse collection of primary hepatocellular carcinomas," *Proceedings of the National Academy of Sciences*, vol. 89, no. 20, pp. 9622–9626, 1992.

[32] T. M. Gottlieb and M. Oren, "p53 in growth control and neoplasia.," *Biochim Biophys Acta*, vol. 1287, pp. 77–102, Jun 1996.

[33] K. W. Kinzler and B. Vogelstein, "Lessons from hereditary colorectal cancer.," *Cell*, vol. 87, pp. 159–170, Oct 1996.

[34] W. S. Mason, C. Aldrich, J. Summers, and J. M. Taylor, "Asymmetric replication of duck hepatitis B virus DNA in liver cells: Free minus-strand DNA.," *Proc Natl Acad Sci U S A*, vol. 79, pp. 3997–4001, Jul 1982.

[35] P. Tiollais, C. Pourcel, and A. Dejean, "The hepatitis B virus.," *Nature*, vol. 317, no. 6037, pp. 489–495, 1985.

[36] R. C. Hirsch, J. E. Lavine, L. J. Chang, H. E. Varmus, and D. Ganem, "Polymerase gene products of hepatitis B viruses are required for genomic RNA packaging as wel as for reverse transcription.," *Nature*, vol. 344, pp. 552–555, Apr 1990.

[37] J. H. Ou, O. Laub, and W. J. Rutter, "Hepatitis B virus gene function: the precore region targets the core antigen to cellular membranes and causes the secretion of the e antigen," *Proceedings of the National Academy of Sciences*, vol. 83, no. 6, pp. 1578–1582, 1986.

[38] T. Kamimura, A. Yoshikawa, F. Ichida, and H. Sasaki, "Electron microscopic studies of Dane particles in hepatocytes with special reference to intracellular development of Dane particles and their relation with HBeAg in serum.," *Hepatology*, vol. 1, no. 5, pp. 392–397, 1981.

[39] A. Alberti, W. H. Gerlich, K. H. Heermann, and P. Pontisso, "Nature and display of hepatitis B virus envelope proteins and the humoral immune response.," *Springer Semin Immunopathol*, vol. 12, no. 1, pp. 5–23, 1990.

[40] D. Fernholz, P. R. Galle, M. Stemler, M. Brunetto, F. Bonino, and H. Will, "Infectious hepatitis B virus variant defective in pre-s2 protein expression in a chronic carrier," *Virology*, vol. 194, no. 1, pp. 137 – 148, 1993.

[41] R. E. Lanford, Y.-H. Kim, H. Lee, L. Notvall, and B. Beames, "Mapping of the hepatitis B virus reverse transcriptase TP and RT domains by transcomplementation for nucleotide priming and by protein-protein interaction," *Journal of Virology*, vol. 73, no. 3, pp. 1885–1893, 1999.

[42] J. Hu, D. O. Toft, and C. Seeger, "Hepadnavirus assembly and reverse transcription require a multicomponent chaperone complex which is incorporated into nucleocapsids.," *EMBO J*, vol. 16, pp. 59–68, Jan 1997.

[43] R. Bartenschlager, M. Junker-Niepmann, and H. Schaller, "The P gene product of hepatitis B virus is required as a structural component for genomic RNA encapsidation.," *J Virol*, vol. 64, pp. 5324–5332, Nov 1990.

[44] Y. Chen and P. L. Marion, "Amino acids essential for RNase H activity of hepadnaviruses are also required for efficient elongation of minus-strand viral DNA.," *J Virol*, vol. 70, pp. 6151–6156, Sep 1996.

[45] S. Kaneko and R. H. Miller, "X-region-specific transcript in mammalian hepatitis B virus-infected liver.," *Journal of Virology*, vol. 62, no. 11, pp. 3979–3984, 1988.

[46] W. T. Guo, J. Wang, G. Tam, T. S. Yen, and J. S. Ou, "Leaky transcription termination produces larger and smaller than genome size hepatitis B virus x gene transcripts.," *Virology*, vol. 181, pp. 630–636, Apr 1991.

[47] E. Pfaff, J. Salfeld, K. Gmelin, H. Schaller, and L. Theilmann, "Synthesis of the x-protein of hepatitis B virus *in vitro* and detection of anti-x antibodies in human sera.," *Virology*, vol. 158, pp. 456–460, Jun 1987.

[48] W. L. Wang, W. T. London, L. Lega, and M. A. Feitelson, "Hbxag in the liver from carrier patients with chronic hepatitis and cirrhosis.," *Hepatology*, vol. 14, pp. 29–37, Jul 1991.

[49] C. M. Kim, K. Koike, I. Saito, T. Miyamura, and G. Jay, "Hbx gene of hepatitis B virus induces liver cancer in transgenic mice.," *Nature*, vol. 351, pp. 317–320, May 1991.

[50] K. Koike, K. Moriya, S. Iino, H. Yotsuyanagi, Y. Endo, T. Miyamura, and K. Kurokawa, "High-level expression of hepatitis B virus hbx gene and hepatocarcinogenesis in transgenic mice," *Hepatology*, vol. 19, no. 4, pp. 810–819, 1994.

[51] T. H. Lee, M. J. Finegold, R. F. Shen, J. L. DeMayo, S. L. Woo, and J. S. Butel, "Hepatitis B virus transactivator x protein is not tumorigenic in transgenic mice.," *J Virol*, vol. 64, pp. 5939–5947, Dec 1990.

[52] M. E. McLaughlin-Drubin and K. Munger, "Viruses associated with human cancer.," *Biochim Biophys Acta*, vol. 1782, pp. 127–150, Mar 2008.

[53] J. Benn and R. J. Schneider, "Hepatitis B virus hbx protein deregulates cell cycle checkpoint controls.," *Proc Natl Acad Sci U S A*, vol. 92, pp. 11215–11219, Nov 1995.

[54] H. Sirma, C. Giannini, K. Poussin, P. Paterlini, D. Kremsdorf, and C. Bréchot, "Hepatitis B virus x mutants, present in hepatocellular carcinoma tissue abrogate both the antiproliferative and transactivation effects of hbx.," *Oncogene*, vol. 18, pp. 4848–4859, Aug 1999.

[55] M. J. B. Tricia L. Gearhart, "The hepatitis B virus X protein modulates hepatocyte proliferation pathways to stimulate viral replication.," *Journal of Virolo*, vol. 84, pp. 2675–2686, Mar 2010.

[56] M. Yamashita and M. Emerman, "Retroviral infection of non-dividing cells: old and new perspectives.," *Virology*, vol. 344, pp. 88–93, Jan 2006.

[57] S. L. McClain, A. J. Clippinger, R. Lizzano, and M. J. Bouchard, "Hepatitis B virus replication is associated with an hbx-dependent mitochondrion-regulated increase in cytosolic calcium levels.," *J Virol*, vol. 81, pp. 12061–12065, Nov 2007.

[58] M. J. Berridge, "Calcium signalling and cell proliferation," *BioEssays*, vol. 17, no. 6, pp. 491–500, 1995.

[59] M. J. Bouchard and R. J. Schneider, "The enigmatic X gene of hepatitis B virus.," *Journal of Virology*, vol. 78, pp. 12725–12734, Dec 2004.

[60] Katsuro and Koike, "Hepatitis B virus x gene is implicated in liver carcinogenesis," *Cancer Letters*, vol. 286, no. 1, pp. 60 – 68, 2009. ¡ce:title¿Hepatocellular - A Worldwide Translational Approach¡/ce:title¿.

[61] A. R. Jilbert, T. T. Wu, J. M. England, P. M. Hall, N. Z. Carp, A. P. O'Connell, and W. S. Mason, "Rapid resolution of duck hepatitis B virus infections occurs after massive hepatocellular involvement.," *J Virol*, vol. 66, pp. 1377–1388, Mar 1992.

[62] K. Kajino, A. R. Jilbert, J. Saputelli, C. E. Aldrich, J. Cullen, and W. S. Mason, "Woodchuck hepatitis virus infections: very rapid recovery after a prolonged viremia and infection of virtually every hepatocyte.," *J Virol*, vol. 68, pp. 5792–5803, Sep 1994.

[63] R. Thimme, S. Wieland, C. Steiger, J. Ghrayeb, K. A. Reimann, R. H. Purcell, and F. V. Chisari, "CD8(+) T cells mediate viral clearance and disease pathogenesis during acute hepatitis B virus infection.," *J Virol*, vol. 77, pp. 68–76, Jan 2003.

[64] J. Y. Wu, Z. Y. Zhou, A. Judd, C. A. Cartwright, and W. S. Robinson, "The hepatitis B virus-encoded transcriptional trans-activator hbx appears to be a novel protein serine/threonine kinase.," *Cell*, vol. 63, pp. 687–695, Nov 1990.

[65] A. Capovilla, *The Role of the Hepatitis B Virus X Protein in Nucleotide Excision Repair*. PhD thesis, University of the Witwatersrand, 2003.

[66] M. A. Romanos, C. A. Scorer, and J. J. Clare, "Foreign gene expression in yeast: a review.," *Yeast*, vol. 8, pp. 423–488, Jun 1992.

[67] S. G. Park, Y. Kim, E. Park, H. M. Ryu, and G. Jung, "Fidelity of hepatitis B virus polymerase.," *Eur J Biochem*, vol. 270, pp. 2929–2936, Jul 2003.

[68] E. Domingo and J. Gomez, "Quasispecies and its impact on viral hepatitis.," *Virus Res*, vol. 127, pp. 131–150, Aug 2007.

[69] S. A. Whalley, J. M. Murray, D. Brown, G. J. Webster, V. C. Emery, G. M. Dusheiko, and A. S. Perelson, "Kinetics of acute hepatitis B virus infection in humans.," *J Exp Med*, vol. 193, pp. 847–854, Apr 2001.

[70] K. M. Weinberger, T. Bauer, S. B$^{\acute{}}$ohm, and W. Jilg, "High genetic variability of the group-specific a-determinant of hepatitis B virus surface antigen (HBsAg) and the corresponding fragment of the viral polymerase in chronic virus carriers lacking detectable HBsAg in serum," *Journal of General Virology*, vol. 81, no. 5, pp. 1165–1174, 2000.

[71] H. Norder, A. M. Courouce, and L. O. Magnius, "Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes.," *Virology*, vol. 198, pp. 489–503, Feb 1994.

[72] C.-J. Chu and A. S. F. Lok, "Clinical significance of hepatitis B virus genotypes.," *Hepatology*, vol. 35, pp. 1274–1276, May 2002.

[73] C.-J. Liu and J.-H. Kao, "Genetic variability of hepatitis B virus and response to antiviral therapy.," *Antivir Ther*, vol. 13, no. 5, pp. 613–624, 2008.

[74] C.-L. Lin and J.-H. Kao, "The clinical implications of hepatitis B virus genotype: Recent advances.," *J Gastroenterol Hepatol*, vol. 26 Suppl 1, pp. 123–130, Jan 2011.

[75] S. Schaefer, “Hepatitis B virus taxonomy and hepatitis B virus genotypes.,” *World J Gastroenterol*, vol. 13, pp. 14–21, Jan 2007.

[76] Z. Yang, *Computational molecular evolution*. Oxford series in ecology and evolution, Oxford University Press, 2006.

[77] T. Yen, “Hepadnaviral x protein:review of recent progress.,” *J Biomed Sci*, vol. 3, pp. 20–30, Jan 1996.

[78] S. Murakami, “Hepatitis B virus X protein: structure, function and biology.,” *Intervirology*, vol. 42, no. 2-3, pp. 81–99, 1999.

[79] J. K. Yee, “A liver-specific enhancer in the core promoter region of human hepatitis B virus.,” *Science*, vol. 246, pp. 658–661, Nov 1989.

[80] Q. Di, J. Summers, J. B. Burch, and W. S. Mason, “Major differences between WHV and HBV in the regulation of transcription.,” *Virology*, vol. 229, pp. 25–35, Mar 1997.

[81] S. Murakami, M. Uchijima, A. Shimoda, S. Kaneko, K. Kobayashi, and N. Hattori, “Hepadnavirus enhancer and its binding proteins.,” *Gastroenterol Jpn*, vol. 25 Suppl 2, pp. 11–19, Sep 1990.

[82] H. Sirma, R. Weil, O. Rosmorduc, S. Urban, A. Israël, D. Kremsdorf, and C. Bréchot, “Cytosol is the prime compartment of hepatitis B virus X protein where it colocalizes with the proteasome.,” *Oncogene*, vol. 16, pp. 2051–2063, Apr 1998.

[83] I. Haviv, M. Shamay, G. Doitsh, and Y. Shaul, “Hepatitis B virus pX targets TFIIB in transcription coactivation.,” *Mol Cell Biol*, vol. 18, pp. 1562–1569, Mar 1998.

[84] M. Doria, N. Klein, R. Lucito, and R. J. Schneider, “The hepatitis B virus HBx protein is a dual specificity cytoplasmic activator of Ras and nuclear activator of transcription factors.,” *EMBO J*, vol. 14, pp. 4747–4757, Oct 1995.

[85] T. Nomura, Y. Lin, D. Dorjsuren, S. Ohno, T. Yamashita, and S. Murakami, “Human hepatitis B virus X protein is detectable in nuclei of transfected cells, and is active for transactivation.,” *Biochim Biophys Acta*, vol. 1453, pp. 330–340, Mar 1999.

[86] F. Henkler, J. Hoare, N. Waseem, R. D. Goldin, M. J. McGarvey, R. Koshy, and I. A. King, “Intracellular localization of the hepatitis B virus HBx protein.,” *J Gen Virol*, vol. 82, pp. 871–882, Apr 2001.

[87] K.-H. Kim and B. L. Seong, “Pro-apoptotic function of HBV X protein is mediated by interaction with c-FLIP and enhancement of death-inducing signal.,” *EMBO J*, vol. 22, pp. 2104–2116, May 2003.

[88] X. W. Wang, M. K. Gibson, W. Vermeulen, H. Yeh, K. Forrester, H. W. Stürzbecher, J. H. Hoeijmakers, and C. C. Harris, “Abrogation of p53-induced apoptosis by the hepatitis B virus X gene.,” *Cancer Research*, vol. 55, pp. 6012–6016, Dec 1995.

[89] P. Chirillo, S. Pagano, G. Natoli, P. L. Puri, V. L. Burgio, C. Balsano, and M. Levrero, “The hepatitis B virus X gene induces p53-mediated programmed cell death.,” *Proc Natl Acad Sci U S A*, vol. 94, pp. 8162–8167, Jul 1997.

[90] F. Su and R. J. Schneider, “Hepatitis B virus HBx protein sensitizes cells to apoptotic killing by tumor necrosis factor alpha.,” *Proc Natl Acad Sci U S A*, vol. 94, pp. 8744–8749, Aug 1997.

[91] F. Zoulim, J. Saputelli, and C. Seeger, “Woodchuck hepatitis virus X protein is required for viral infection *in vivo*.,” *J Virol*, vol. 68, pp. 2026–2030, Mar 1994.

[92] H. S. Chen, S. Kaneko, R. Girones, R. W. Anderson, W. E. Hornbuckle, B. C. Tennant, P. J. Cote, J. L. Gerin, R. H. Purcell, and R. H. Miller, “The woodchuck hepatitis virus X gene is important for establishment of virus infection in woodchucks.,” *J Virol*, vol. 67, pp. 1218–1226, Mar 1993.

[93] V. V. Keasler, A. J. Hodgson, C. R. Madden, and B. L. Slagle, “Hepatitis B virus HBx protein localized to the nucleus restores HBx-deficient virus replication in HepG2 cells and *in vivo* in hydrodynamically-injected mice.,” *Virology*, vol. 390, pp. 122–129, Jul 2009.

[94] M.-Y. Cha, D.-K. Ryu, H.-S. Jung, H.-E. Chang, and W.-S. Ryu, “Stimulation of hepatitis B virus genome

replication by HBx is linked to both nuclear and cytoplasmic HBx expression.," *J Gen Virol*, vol. 90, pp. 978–986, Apr 2009.

[95] S. Jameel, A. Siddiqui, H. F. Maguire, and K. V. Rao, "Hepatitis B virus X protein produced in *Escherichia coli* is biologically functional.," *Journal of Virology*, vol. 64, pp. 3963–3966, Aug 1990.

[96] A. Gupta, T. K. Mal, N. Jayasuryan, and V. S. Chauhan, "Assignment of disulphide bonds in the X protein (HBx) of hepatitis B virus.," *Biochem Biophys Res Commun*, vol. 212, pp. 919–924, Jul 1995.

[97] I. Marczinovits, C. Somogyi, A. Patthy, P. Németh, and J. Molnár, "An alternative purification protocol for producing hepatitis B virus X antigen on a preparative scale in *Escherichia coli*.," *J Biotechnol*, vol. 56, pp. 81–88, Aug 1997.

[98] S. Urban, E. Hildt, C. Eckerskorn, H. Sirma, A. Kekulé, and P. H. Hofschneider, "Isolation and molecular characterization of hepatitis B virus X-protein from a baculovirus expression system.," *Hepatology*, vol. 26, pp. 1045–1053, Oct 1997.

[99] Z. Hu, Z. Zhang, E. Doo, O. Coux, A. L. Goldberg, and T. J. Liang, "Hepatitis B virus X protein is both a substrate and a potential inhibitor of the proteasome complex.," *Journal of Virology*, vol. 73, pp. 7231–7240, Sep 1999.

[100] S. H. Lwa and W. N. Chen, "Hepatitis B virus X protein interacts with β 5 subunit of heterotrimeric guanine nucleotide binding protein.," *Virology Journal*, vol. 2, p. 76, 2005.

[101] M. R. Banki, L. Feng, and D. W. Wood, "Simple bioseparations using self-cleaving elastin-like polypeptide tags.," *Nat Methods*, vol. 2, pp. 659–661, Sep 2005.

[102] M. S. Mufamadi, "Inhibition of hepatitis B virus subgenotype A1 replication using activators of RNA interference," Master's thesis, 2009.

[103] M. Kozak, "An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs.," *Nucleic Acids Res*, vol. 15, pp. 8125–8148, Oct 1987.

[104] M. A. Sheff and K. S. Thorn, "Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*.," *Yeast*, vol. 21, pp. 661–670, Jun 2004.

[105] A. Capovilla and P. Arbuthnot, "Hepatitis B virus X protein does not influence essential steps of nucleotide excision repair effected by human liver extracts.," *Biochem Biophys Res Commun*, vol. 312, pp. 806–810, Dec 2003.

[106] B. A. Fong, W.-Y. Wu, and D. W. Wood, "Optimization of elp-intein mediated protein purification by salt substitution.," *Protein Expr Purif*, vol. 66, pp. 198–202, Aug 2009.

[107] W.-Y. Wu, C. Mee, F. Califano, R. Banki, and D. W. Wood, "Recombinant protein purification by self-cleaving aggregation tag.," *Nat Protoc*, vol. 1, no. 5, pp. 2257–2262, 2006.

[108] D. Huang, P. R. Gore, and E. V. Shusta, "Increasing yeast secretion of heterologous proteins by regulating expression rates and post-secretory loss.," *Biotechnol Bioeng*, vol. 101, pp. 1264–1275, Dec 2008.

[109] S. M. Spinola and J. G. Cannon, "Different blocking agents cause variation in the immunologic detection of proteins transferred to nitrocellulose membranes.," *J Immunol Methods*, vol. 81, pp. 161–165, Jul 1985.

[110] E. T. Thean and B. H. Toh, "Western immunoblotting: temperature-dependent reduction in background staining.," *Anal Biochem*, vol. 177, pp. 256–258, Mar 1989.

[111] M. Chevallet, S. Luche, and T. Rabilloud, "Silver staining of proteins in polyacrylamide gels.," *Nat Protoc*, vol. 1, no. 4, pp. 1852–1858, 2006.

[112] A. L. Atkin, "Preparation of yeast cells for confocal microscopy," vol. 122, December 1998.

[113] S. Nadin-Davis and V. A. Mezl, "Optimization of the ethanol precipitation of RNA from formamide containing solutions.," *Prep Biochem*, vol. 12, no. 1, pp. 49–56, 1982.

[114] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "Globplot: Exploring protein sequences for globularity and disorder.," *Nucleic Acids Res*, vol. 31, pp. 3701–3708, Jul 2003.

[115] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.," *Bioinformatics*, vol. 21, pp. 3369–3376, Aug 2005.

[116] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, "Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.," *Bioinformatics*, vol. 21, pp. 3433–3434, Aug 2005.

[117] T. Ishida and K. Kinoshita, "Prdos: prediction of disordered protein regions from amino acid sequence.," *Nucleic Acids Res*, vol. 35, pp. W460–W464, Jul 2007.

[118] "Protein disorder predictors." http://www.disprot.org/, 2011.

[119] "EPESTFIND." http://emboss.bioinformatics.nl/.

[120] M. B. Dana Reichmann, Shmuel Pietrokovski, "Online destruction box motif (dbox) prediction tool." `http://bioinfo.weizmann.ac.il/~danag/d-box/main.html`.

[121] T. P. S. John and R. W. Davis, "The organization and transcription of the galactose gene cluster of saccharomyces.," *J Mol Biol*, vol. 152, pp. 285–315, Oct 1981.

[122] M. Johnston, "A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces cerevisiae*.," *Microbiol Rev*, vol. 51, pp. 458–476, Dec 1987.

[123] J. R. Shuster, "Regulated transcriptional systems for the production of proteins in yeast: regulation by carbon source.," *Biotechnology*, vol. 13, pp. 83–108, 1989.

[124] V. L. Price, W. E. Taylor, W. Clevenger, M. Worthington, and E. T. Young, "Expression of heterologous proteins in *Saccharomyces cerevisiae* using the ADH2 promoter.," *Methods Enzymol*, vol. 185, pp. 308–318, 1990.

[125] M. Rechsteiner and S. W. Rogers, "Pest sequences and regulation by proteolysis.," *Trends Biochem Sci*, vol. 21, pp. 267–271, Jul 1996.

[126] S. Rogers, R. Wells, and M. Rechsteiner, "Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis.," *Science*, vol. 234, pp. 364–368, Oct 1986.

[127] K. Guruprasad, B. V. Reddy, and M. W. Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence.," *Protein Engineering*, vol. 4, pp. 155–161, Dec 1990.

[128] D. Liu, L. Zou, W. Li, L. Wang, and Y. Wu, "High-level expression and large-scale preparation of soluble HBx antigen from *Escherichia coli*.," *Biotechnology and Applied Biochemistry*, vol. 54, pp. 141–147, Nov 2009.

[129] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast.," *Mol Cell Biol*, vol. 19, pp. 1720–1730, Mar 1999.

[130] A. Mehra, K. H. Lee, and V. Hatzimanikatis, "Insights into the relation between mRNA and protein expression patterns: I. theoretical considerations.," *Biotechnology and Bioengineering*, vol. 84, pp. 822–833, Dec 2003.

[131] P. S. Lee, L. B. Shaw, L. H. Choe, A. Mehra, V. Hatzimanikatis, and K. H. Lee, "Insights into the relation between mRNA and protein expression patterns: Ii. experimental observations in *Escherichia coli*.," *Biotechnol Bioeng*, vol. 84, pp. 834–841, Dec 2003.

[132] N. A. Burgess-Brown, S. Sharma, F. Sobott, C. Loenarz, U. Oppermann, and O. Gileadi, "Codon optimization can improve expression of human genes in *Escherichia coli*: A multi-gene study.," *Protein Expr Purif*, vol. 59, pp. 94–102, May 2008.

[133] L. Kotula and P. J. Curtis, "Evaluation of foreign gene codon optimization in yeast: expression of a

mouse IG kappa chain.," *Biotechnology (N Y)*, vol. 9, pp. 1386–1389, Dec 1991.

[134] A. Yadava and C. F. Ockenhouse, "Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems.," *Infect Immun*, vol. 71, pp. 4961–4969, Sep 2003.

[135] T. T. Yang, L. Cheng, and S. R. Kain, "Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein.," *Nucleic Acids Res*, vol. 24, pp. 4592–4593, Nov 1996.

[136] A. Basu, W. N. Chen, and S. S. J. Leong, "A rational design for hepatitis B virus X protein refolding and bioprocess development guided by second virial coefficient studies.," *Appl Microbiol Biotechnol*, vol. 90, pp. 181–191, Apr 2011.

[137] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes.," *Genome Inform Ser Workshop Genome Inform*, vol. 11, pp. 161–171, 2000.

[138] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins.," *J Mol Biol*, vol. 323, pp. 573–584, Oct 2002.

[139] Y. Minezaki, K. Homma, A. R. Kinjo, and K. Nishikawa, "Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation.," *J Mol Biol*, vol. 359, pp. 1137–1149, Jun 2006.

[140] B. Xue, R. W. Williams, C. J. Oldfield, G. K.-M. Goh, A. K. Dunker, and V. N. Uversky, "Viral disorder or disordered viruses: do viral proteins possess unique features?," *Protein Pept Lett*, vol. 17, pp. 932–951, Aug 2010.

[141] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright, "Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity.," *Proc Natl Acad Sci U S A*, vol. 93, pp. 11504–11509, Oct 1996.

[142] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26 – 59, 2001.

[143] P. Tompa, "Intrinsically unstructured proteins.," *Trends Biochem Sci*, vol. 27, pp. 527–533, Oct 2002.

[144] P. Tompa, J. Prilusky, I. Silman, and J. L. Sussman, "Structural disorder serves as a weak signal for intracellular protein degradation," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, no. 2, pp. 903–909, 2008.

[145] M. Yazawa, F. Matsuzawa, and K. Yagi, "Inter-domain interaction and the structural flexibility of calmodulin in the connecting region of the terminal two domains.," *J Biochem*, vol. 107, pp. 287–291, Feb 1990.

[146] B. Chakravarthy, P. Morley, and J. Whitfield, "Ca_{2+}-calmodulin and protein kinase cs: a hypothetical synthesis of their conflicting convergences on shared substrate domains.," *Trends Neurosci*, vol. 22, pp. 12–16, Jan 1999.

[147] C.-W. Liu, M. J. Corboy, G. N. DeMartino, and P. J. Thomas, "Endoproteolytic activity of the proteasome.," *Science*, vol. 299, pp. 408–411, Jan 2003.

[148] G. Asher, N. Reuven, and Y. Shaul, "20S proteasomes and protein degradation "by default"," *Bioessays*, vol. 28, pp. 844–849, Aug 2006.

[149] P. Tsvetkov, G. Asher, A. Paz, N. Reuven, J. L. Sussman, I. Silman, and Y. Shaul, "Operational definition of intrinsically unstructured protein sequences based on susceptibility to the 20s proteasome.," *Proteins*, vol. 70, pp. 1357–1366, Mar 2008.

[150] S. Prakash, L. Tian, K. S. Ratliff, R. E. Lehotzky, and A. Matouschek, "An unstructured initiation site is required for efficient proteasome-mediated degradation.," *Nat Struct Mol Biol*, vol. 11, pp. 830–837, Sep 2004.

[151] J. Zhao, C. Wang, J. Wang, X. Yang, N. Diao, Q. Li, W. Wang, L. Xian, Z. Fang, and L. Yu, "E3 ubiquitin ligase Siah-1 facilitates poly-ubiquitylation and proteasomal degradation of the hepatitis B viral X protein," *FEBS Letters*, vol. In Press, Uncorrected Proof, pp. –, 2011.

[152] M. Fuxreiter, P. Tompa, and I. Simon, "Local structural disorder imparts plasticity on linear motifs.," *Bioinformatics*, vol. 23, pp. 950–956, Apr 2007.

[153] G. P. Singh, M. Ganapathi, K. S. Sandhu, and D. Dash, "Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes.," *Proteins*, vol. 62, pp. 309–315, Feb 2006.

[154] J. F. Dice, "Molecular determinants of protein half-lives in eukaryotic cells.," *FASEB J*, vol. 1, pp. 349–357, Nov 1987.

[155] H. Tang, L. Delgermaa, F. Huang, N. Oishi, L. Liu, F. He, L. Zhao, and S. Murakami, "The transcriptional transactivation function of HBx protein is important for its augmentation role in hepatitis B virus replication.," *J Virol*, vol. 79, pp. 5548–5556, May 2005.

[156] S. Murakami, J. H. Cheong, and S. Kaneko, "Human hepatitis virus X gene encodes a regulatory domain that represses transactivation of X protein.," *J Biol Chem*, vol. 269, pp. 15118–15123, May 1994.

[157] N. Blow, "Antibodies: The generation game," *Nature*, vol. 447, pp. 741–744, June 2007.

[158] B. Friguet, L. Djavadi-Ohaniance, J. Pages, A. Bussard, and M. Goldberg, "A convenient enzyme-linked immunosorbent assay for testing whether monoclonal antibodies recognize the same antigenic site. application to hybridomas specific for the β_2-subunit of *Escherichia coli* tryptophan synthase.," *J Immunol Methods*, vol. 60, pp. 351–358, Jun 1983.

[159] L. Djavadi-Ohaniance, B. Friguet, and M. E. Goldberg, "Structural and functional influence of enzyme-antibody interactions: effects of eight different monoclonal antibodies on the enzymatic activity of *Escherichia coli* tryptophan synthase.," *Biochemistry*, vol. 23, pp. 97–104, Jan 1984.

[160] B. Friguet, L. Djavadi-Ohaniance, and M. E. Goldberg, "Some monoclonal antibodies raised with a native protein bind preferentially to the denatured antigen.," *Mol Immunol*, vol. 21, pp. 673–677, Jul 1984.

[161] H. C. Vaidya, D. N. Dietzler, and J. H. Ladenson, "Inadequacy of traditional ELISA for screening hybridoma supernatants for murine monoclonal antibodies.," *Hybridoma*, vol. 4, no. 3, pp. 271–276, 1985.

[162] M. L. Overall, S. Marzuki, and P. J. Hertzog, "Comparison of different ELISAs for the detection of monoclonal antibodies to human interferon-alpha. implications for antibody screening.," *J Immunol Methods*, vol. 119, pp. 27–33, Apr 1989.

[163] R. B. Kapust and D. S. Waugh, "*Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.," *Protein Sci*, vol. 8, pp. 1668–1674, Aug 1999.

[164] J. D. Fox, R. B. Kapust, and D. S. Waugh, "Single amino acid substitutions on the surface of *Escherichia coli* maltose-binding protein can have a profound impact on the solubility of fusion proteins.," *Protein Sci*, vol. 10, pp. 622–630, Mar 2001.

[165] G. Richarme and T. D. Caldas, "Chaperone properties of the bacterial periplasmic substrate-binding proteins.," *J Biol Chem*, vol. 272, pp. 15607–15612, Jun 1997.

[166] D. Sachdev and J. M. Chirgwin, "Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin.," *Protein Expr Purif*, vol. 12, pp. 122–132, Feb 1998.

[167] S. Nallamsetty and D. S. Waugh, "Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners.," *Protein Expression and Purification*, vol. 45, pp. 175–182, Jan 2006.

[168] K. Lundstrom, "Applications of the SFV system for studies on receptors and ion channels.," *Focus*, vol. 18, pp. 53–56, 1996.

[169] J. R. Rodriguez-Madoz, J. Prieto, and C. Smerdou, "Semliki Forest virus vectors engineered to express higher IL-12 levels induce efficient elimination of murine colon adenocarcinomas.," *Molecular Therapy*, vol. 12, pp. 153–163, Jul 2005.

[170] E. Casales, J. R. Rodriguez-Madoz, M. Ruiz-Guillen, N. Razquin, Y. Cuevas, J. Prieto, and C. Smerdou, "Development of a new noncytopathic Semliki Forest virus vector providing high expression levels and stability.," *Virology*, vol. 376, pp. 242–251, Jun 2008.

[171] V. O. R. B. u. J. A. P. P. H. G. Tatyana Kozlovska, Anna Zajakina, "Synthesis of all hepatitis B structural proteins in the Semliki Forest virus expression system," *Acta Universitatis Latviensis, Biology*, vol. 674, pp. 39–51, 2004.

[172] J. H. Ou, H. Bao, C. Shih, and S. M. Tahara, "Preferred translation of human hepatitis B virus polymerase from core protein- but not from precore protein-specific transcript.," *Journal of Virology*, vol. 64, no. 9, pp. 4578–4581, 1990.

[173] C. Smerdou and P. Liljeström, "Two-helper RNA system for production of recombinant Semliki Forest virus particles.," *Journal of Virology*, vol. 73, pp. 1092–1098, Feb 1999.

[174] S. C. K. Ryan B. Jensen, Aura Carreira, "Purified human BRCA2 stimulates RAD51-mediated recombination.," *Nature*, vol. 467, pp. 678–683, Oct 2010.

[175] E. M. Sjöberg, M. Suomalainen, and H. Garoff, "A significantly improved Semliki Forest virus expression system based on translation enhancer segments from the viral capsid gene.," *Biotechnology (N Y)*, vol. 12, pp. 1127–1131, Nov 1994.

[176] Y. Mahé, N. Mukaida, K. Kuno, M. Akiyama, N. Ikeda, K. Matsushima, and S. Murakami, "Hepatitis B virus X protein transactivates human interleukin-8 gene through acting on nuclear factor kB and CCAAT/enhancer-binding protein-like cis-elements.," *Journal of Biological Chemistry*, vol. 266, no. 21, pp. 13759–13763, 1991.

[177] W. Rasband, "Imagej." U.S. National Institutes of Health, Bethesda, Maryland, USA 1997-2011.

[178] M. Green and P. M. Loewenstein, "Autonomous functional domains of chemically synthesized human immunodeficiency virus tat trans-activator protein.," *Cell*, vol. 55, pp. 1179–1188, Dec 1988.

[179] A. D. Frankel and C. O. Pabo, "Cellular uptake of the tat protein from human immunodeficiency virus.," *Cell*, vol. 55, pp. 1189–1193, Dec 1988.

[180] C. Balsano, O. Billet, M. Bennoun, C. Cavard, A. Zider, G. Grimber, G. Natoli, P. Briand, and M. Levrero, "The hepatitis B virus X gene product transactivates the HIV-LTR *in vivo*.," *Arch Virol Suppl*, vol. 8, pp. 63–71, 1993.

[181] G. Griffiths, G. Warren, P. Quinn, O. Mathieu-Costello, and H. Hoppeler, "Density of newly synthesized plasma membrane proteins in intracellular membranes. i. stereological studies.," *Journal of Cellular Biology*, vol. 98, pp. 2133–2141, Jun 1984.

[182] S. Raghava, S. Aquil, S. Bhattacharyya, R. Varadarajan, and M. N. Gupta, "Strategy for purifying maltose binding protein fusion proteins by affinity precipitation.," *J Chromatogr A*, vol. 1194, pp. 90–95, Jun 2008.

[183] J. G. Marblestone, S. C. Edavettal, Y. Lim, P. Lim, X. Zuo, and T. R. Butt, "Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO.," *Protein Science*, vol. 15, pp. 182–189, Jan 2006.

[184] F. Wurm and A. Bernard, "Large-scale transient expression in mammalian cells for recombinant protein production.," *Curr Opin Biotechnol*, vol. 10, pp. 156–159, Apr 1999.

[185] K. Lundstrom, A. Abenavoli, A. Malgaroli, and M. U. Ehrengruber, "Novel Semliki Forest virus vectors with reduced cytotoxicity and temperature sensitivity for long-term enhancement of transgene expression.," *Molecular Therapy*, vol. 7, pp. 202–209, Feb 2003.

[186] E. Casales, A. Aranda, J. I. Quetglas, M. Ruiz-Guillen, J. R. Rodriguez-Madoz, J. Prieto, and C. Smerdou, "A novel system for the production

of high levels of functional human therapeutic proteins in stable cells with a Semliki Forest virus noncytopathic vector.," *Nature Biotechnology*, vol. 27, pp. 138–148, May 2010.

[187] P. Melancon and H. Garoff, "Processing of the Semliki Forest virus structural polyprotein: role of the capsid protease.," *Journal of Virology*, vol. 61, pp. 1301–1309, May 1987.

[188] C. S. Hahn and J. H. Strauss, "Site-directed mutagenesis of the proposed catalytic amino acids of the Sindbis virus capsid protein autoprotease.," *J Virol*, vol. 64, pp. 3069–3073, Jun 1990.

[189] G. J. Atkins, "Establishment of persistent infection in BHK-21 cells by temperature-sensitive mutants of Sindbis virus.," *J Gen Virol*, vol. 45, pp. 201–207, Oct 1979.

[190] P. Grassé, *Evolution of living organisms: evidence for a new theory of transformation*. Academic Press, 1977.

[191] B. G. Barrell, G. M. Air, and C. A. Hutchison, "Overlapping genes in bacteriophage ϕx174," *Nature*, vol. 264, pp. 34–41, Nov. 1976.

[192] P. K. Keese and A. Gibbs, "Origins of genes: "big bang" or continuous creation?," *Proc Natl Acad Sci U S A*, vol. 89, pp. 9489–9493, Oct 1992.

[193] N. Chirico, A. Vianelli, and R. Belshaw, "Why genes overlap in viruses.," *Proc Biol Sci*, vol. 277, pp. 3809–3817, Dec 2010.

[194] D. C. Krakauer, "Stability and evolution of overlapping genes.," *Evolution*, vol. 54, pp. 731–739, Jun 2000.

[195] O. Peleg, V. Kirzhner, E. Trifonov, and A. Bolshoy, "Overlapping messages and survivability.," *J Mol Evol*, vol. 59, pp. 520–527, Oct 2004.

[196] C. Hannoun, P. Horal, and M. Lindh, "Long-term mutation rates in the hepatitis b virus genome," *Journal of General Virology*, vol. 81, no. 1, pp. 75–83, 2000.

[197] D. C. Krakauer and J. B. Plotkin, "Redundancy, antiredundancy, and the robustness of genomes," *Proceedings of the National Academy of Sciences*, vol. 99, no. 3, pp. 1405–1409, 2002.

[198] C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, and D. Karlin, "Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation.," *J Virol*, vol. 83, pp. 10719–10736, Oct 2009.

[199] R. H. Miller, S. Kaneko, C. T. Chung, R. Girones, and R. H. Purcell, "Compact organization of the hepatitis B virus genome.," *Hepatology*, vol. 9, pp. 322–327, Feb 1989.

[200] H. L. Zaaijer, F. J. van Hemert, M. H. Koppelman, and V. V. Lukashov, "Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus.," *Journal of General Virology*, vol. 88, pp. 2137–2143, Aug 2007.

[201] T. de Oliveira, K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg, A. M. J. Wensing, D. A. van de Vijver, C. A. Boucher, R. Camacho, and A.-M. Vandamme, "An automated genotyping system for analysis of HIV-1 and other microbial sequences.," *Bioinformatics*, vol. 21, pp. 3797–3800, Oct 2005.

[202] L. C. J. Alcantara, S. Cassol, P. Libin, K. Deforche, O. G. Pybus, M. V. Ranst, B. G. ao Castro, A.-M. Vandamme, and T. de Oliveira, "A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences.," *Nucleic Acids Res*, vol. 37, pp. W634–W642, Jul 2009.

[203] P. Rice, I. Longden, and A. Bleasby, "Emboss: the european molecular biology open software suite.," *Trends Genet*, vol. 16, pp. 276–277, Jun 2000.

[204] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput.," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[205] O. Gascuel, "Bionj: an improved version of the NJ algorithm based on a simple model of sequence data.," *Mol Biol Evol*, vol. 14, pp. 685–695, Jul 1997.

[206] M. Gouy, S. Guindon, and O. Gascuel, "Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.," *Mol Biol Evol*, vol. 27, pp. 221–224, Feb 2010.

[207] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood.," *Comput Appl Biosci*, vol. 13, pp. 555–556, Oct 1997.

[208] Z. Yang, "PAML 4: Phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[209] M. Anisimova, J. P. Bielawski, and Z. Yang, "Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution," *Molecular Biology and Evolution*, vol. 18, no. 8, pp. 1585–1592, 2001.

[210] M. Anisimova, J. P. Bielawski, and Z. Yang, "Accuracy and power of bayes prediction of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 950–958, 2002.

[211] Z. Yang, W. S. W. Wong, and R. Nielsen, "Bayes empirical bayes inference of amino acid sites under positive selection.," *Mol Biol Evol*, vol. 22, pp. 1107–1118, Apr 2005.

[212] T. A. Hall, "Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[213] D. T. Pride, "Swaap - sliding windows alignment analysis program," 2000.

[214] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.," *Molecular Biology and Evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[215] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.," *Genetics*, vol. 148, pp. 929–936, Mar 1998.

[216] Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites.," *Genetics*, vol. 155, pp. 431–449, May 2000.

[217] A. Anisimova, *Detecting positive selection in protein-coding genes*. PhD thesis, University College London, 2003.

[218] Yang and Bielawski, "Statistical methods for detecting molecular adaptation.," *Trends Ecol Evol*, vol. 15, pp. 496–503, Dec 2000.

[219] L. D. Hurst, "The Ka/Ks ratio: diagnosing the form of sequence evolution," *Trends in Genetics*, vol. 18, no. 9, pp. 486 – 487, 2002.

[220] S. Kryazhimskiy and J. B. Plotkin, "The population genetics of dN/dS.," *PLoS Genet*, vol. 4, p. e1000304, Dec 2008.

[221] J. Zhang, R. Nielsen, and Z. Yang, "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.," *Mol Biol Evol*, vol. 22, pp. 2472–2479, Dec 2005.

[222] Z. Yang, R. Nielsen, and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution.," *Mol Biol Evol*, vol. 15, pp. 1600–1611, Dec 1998.

[223] J. Zhang, "Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes.," *J Mol Evol*, vol. 50, pp. 56–68, Jan 2000.

[224] I. B. Rogozin, A. N. Spiridonov, A. V. Sorokin, Y. I. Wolf, I. K. Jordan, R. L. Tatusov, and E. V. Koonin, "Purifying and directional selection in overlapping prokaryotic genes.," *Trends Genet*, vol. 18, pp. 228–232, May 2002.

[225] D. B. Searls, "Pharmacophylogenomics: genes, evolution and drug targets.," *Nat Rev Drug Discov*, vol. 2, pp. 613–623, Aug 2003.

[226] P. M. Durand, K. Naidoo, and T. L. Coetzer, "Evolutionary patterning: a novel approach to the identification of potential drug target sites in *Plasmodium falciparum*.," *PLoS One*, vol. 3, no. 11, p. e3685, 2008.

[227] P. Valenzuela, A. Medina, W. J. Rutter, G. Ammerer, and B. D. Hall, "Synthesis and assembly of hepatitis B virus surface antigen particles in yeast.," *Nature*, vol. 298, pp. 347–350, Jul 1982.

[228] A. Miyanohara, T. Imamura, M. Araki, K. Sugawara, N. Ohtomo, and K. Matsubara, "Expression of hepatitis B virus core antigen gene in *Saccharomyces cerevisiae*: synthesis of two polypeptides translated from different initiation codons.," *J Virol*, vol. 59, pp. 176–180, Jul 1986.

[229] I. Qadri and A. Siddiqui, "Expression of hepatitis B virus polymerase in Ty1-his3AI retroelement of *Saccharomyces cerevisiae*.," *Journal of Biological Chemistry*, vol. 274, pp. 31359–31365, Oct 1999.

[230] C. J. Jeffery, "Moonlighting proteins," *Trends Biochem Sci*, vol. 24, pp. 8–11, Jan 1999.

[231] P. S. Low, P. Rathinavelu, and M. L. Harrison, "Regulation of glycolysis via reversible enzyme binding to the membrane protein, band 3.," *J Biol Chem*, vol. 268, pp. 14627–14631, Jul 1993.

[232] C. J. Jeffery, "Moonlighting proteins - an update," *Mol Biosyst*, vol. 5, pp. 345–350, Apr 2009.

[233] P. O. de Spicer and S. Maloy, "Puta protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator.," *Proc Natl Acad Sci U S A*, vol. 90, pp. 4295–4298, May 1993.

[234] S. Soker, S. Takashima, H. Q. Miao, G. Neufeld, and M. Klagsbrun, "Neuropilin-1 is expressed by endothelial and tumor cells as an isoform-specific receptor for vascular endothelial growth factor.," *Cell*, vol. 92, pp. 735–745, Mar 1998.

[235] A. L. Kolodkin, D. V. Levengood, E. G. Rowe, Y. T. Tai, R. J. Giger, and D. D. Ginty, "Neuropilin is a semaphorin iii receptor.," *Cell*, vol. 90, pp. 753–762, Aug 1997.

[236] K. Meyer-Siegler, D. J. Mauro, G. Seal, J. Wurzer, J. K. deRiel, and M. A. Sirover, "A human nuclear uracil DNA glycosylase is the 37-kda subunit of glyceraldehyde-3-phosphate dehydrogenase.," *Proc Natl Acad Sci U S A*, vol. 88, pp. 8460–8464, Oct 1991.

[237] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu, "Tight regulation of unstructured proteins: from transcript synthesis to protein degradation.," *Science*, vol. 322, pp. 1365–1368, Nov 2008.

[238] N. A. Sallee, G. M. Rivera, J. E. Dueber, D. Vasilescu, R. D. Mullins, B. J. Mayer, and W. A. Lim, "The pathogen protein espf(u) hijacks actin polymerization using mimicry and multivalency.," *Nature*, vol. 454, pp. 1005–1008, Aug 2008.

[239] H.-C. Cheng, B. M. Skehan, K. G. Campellone, J. M. Leong, and M. K. Rosen, "Structural mechanism of WASP activation by the enterohaemorrhagic *E. coli* effector EspF(U).," *Nature*, vol. 454, pp. 1009–1013, Aug 2008.

[240] D. M. A. Gendoo, M. M. El-Hefnawi, M. Werner, and R. Siam, "Correlating novel variable and conserved motifs in the hemagglutinin protein with significant biological functions.," *Virol J*, vol. 5, p. 91, 2008.

[241] O. Aitio, M. Hellman, A. Kazlauskas, D. F. Vingadassalom, J. M. Leong, K. Saksela, and P. Permi, "Recognition of tandem PxxP motifs as a unique Src homology 3-binding mode triggers pathogen-driven actin assembly.," *Proc Natl Acad Sci U S A*, vol. 107, pp. 21743–21748, Dec 2010.

[242] A. S. Selyunin, S. E. Sutton, B. A. Weigele, L. E. Reddick, R. C. Orchard, S. M. Bresson, D. R. Tomchick, and N. M. Alto, "The assembly of a GTPase-kinase signalling complex by a bacterial catalytic scaffold.," *Nature*, vol. 469, pp. 107–111, Jan 2011.

[243] D. Kalderon, B. L. Roberts, W. D. Richardson, and A. E. Smith, "A short amino acid sequence able to specify nuclear location.," *Cell*, vol. 39, pp. 499–509, Dec 1984.

[244] R. E. Jones, R. J. Wegrzyn, D. R. Patrick, N. L. Balishin, G. A. Vuocolo, M. W. Riemen, D. Defeo-Jones, V. M. Garsky, D. C. Heimbrook, and A. Oliff, "Identification of HPV-16 E7 peptides that are potent antagonists of E7 binding to the retinoblastoma suppressor protein.," *J Biol Chem*, vol. 265, pp. 12782–12785, Aug 1990.

[245] H. G. Göttlinger, T. Dorfman, J. G. Sodroski, and W. A. Haseltine, "Effect of mutations affecting the p6

gag protein on human immunodeficiency virus particle release.," *Proc Natl Acad Sci U S A*, vol. 88, pp. 3195–3199, Apr 1991.

[246] J. M. Boyd, T. Subramanian, U. Schaeper, M. L. Regina, S. Bayley, and G. Chinnadurai, "A region in the c-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis.," *EMBO Journal*, vol. 12, pp. 469–478, Feb 1993.

[247] T. Vavouri, J. I. Semple, R. Garcia-Verdugo, and B. Lehner, "Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity.," *Cell*, vol. 138, pp. 198–208, Jul 2009.

[248] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Trave, and T. J. Gibson, "Understanding eukaryotic linear motifs and their role in cell signaling and regulation.," *Front Biosci*, vol. 13, pp. 6580–6603, 2008.

[249] R. B. Jones, A. Gordus, J. A. Krall, and G. MacBeath, "A quantitative protein interaction network for the ErbB receptors using protein microarrays.," *Nature*, vol. 439, pp. 168–174, Jan 2006.

[250] N. E. Buchler and M. Louis, "Molecular titration and ultrasensitivity in regulatory networks," *Journal of Molecular Biology*, vol. 384, no. 5, pp. 1106 – 1119, 2008.

[251] M. S. Cortese, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder in scaffold proteins: getting more from less.," *Prog Biophys Mol Biol*, vol. 98, pp. 85–106, Sep 2008.

[252] J. Diao, A. A. Khine, F. Sarangi, E. Hsu, C. Iorio, L. A. Tibbles, J. R. Woodgett, J. Penninger, and C. D. Richardson, "X protein of hepatitis B virus inhibits Fas-mediated apoptosis and is associated with up-regulation of the SAPK/JNK pathway.," *J Biol Chem*, vol. 276, pp. 8328–8340, Mar 2001.

[253] T. Li, E. I. Robert, P. C. van Breugel, M. Strubin, and N. Zheng, "A promiscuous alpha-helical motif anchors viral hijackers and substrate receptors to the CUL4-DDB1 ubiquitin ligase machinery.," *Nat Struct Mol Biol*, vol. 17, pp. 105–111, Jan 2010.

[254] T. L. Tan, Z. Feng, Y. W. Lu, V. Chan, and W. N. Chen, "Adhesion contact kinetics of HepG2 cells during hepatitis B virus replication: Involvement of SH3-binding motif in hbx.," *Biochim Biophys Acta*, vol. 1762, pp. 755–766, Aug 2006.

[255] M. Forgues, A. J. Marrogi, E. A. Spillare, C. G. Wu, Q. Yang, M. Yoshida, and X. W. Wang, "Interaction of the hepatitis B virus x protein with the Crm1-dependent nuclear export pathway.," *J Biol Chem*, vol. 276, pp. 22797–22803, Jun 2001.

[256] Q. Su, C. H. Schröder, W. J. Hofmann, G. Otto, R. Pichlmayr, and P. Bannasch, "Expression of hepatitis B virus X protein in HBV-infected human livers and hepatocellular carcinomas.," *Hepatology*, vol. 27, pp. 1109–1120, Apr 1998.

[257] T. L. Deans, C. R. Cantor, and J. J. Collins, "A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells.," *Cell*, vol. 130, pp. 363–372, Jul 2007.

[258] A. C. Joerger and A. R. Fersht, "Structure-function-rescue: the diverse nature of common p53 cancer mutants.," *Oncogene*, vol. 26, pp. 2226–2242, Apr 2007.

[259] S. Hosono, M. J. Chou, C. S. Lee, and C. Shih, "Infrequent mutation of p53 gene in hepatitis B virus positive primary hepatocellular carcinomas.," *Oncogene*, vol. 8, pp. 491–496, Feb 1993.

[260] H. Ueda, S. J. Ullrich, J. D. Gangemi, C. A. Kappel, L. Ngo, M. A. Feitelson, and G. Jay, "Functional inactivation but not structural mutation of p53 causes liver cancer.," *Nature Genetics*, vol. 9, pp. 41–47, Jan 1995.

[261] M. S. Greenblatt, M. A. Feitelson, M. Zhu, W. P. Bennett, J. A. Welsh, R. Jones, A. Borkowski, and C. C. Harris, "Integrity of p53 in hepatitis B X antigen-positive and -negative hepatocellular carcinomas," *Cancer Research*, vol. 57, no. 3, pp. 426–432, 1997.

[262] O. Terradillos, T. Pollicino, H. Lecoeur, M. Tripodi, M. L. Gougeon, P. Tiollais, and M. A. Buendia, "p53-independent apoptotic effects of the hepatitis B virus HBx protein *in vivo* and *in vitro*.," *Oncogene*, vol. 17, pp. 2115–2123, Oct 1998.

[263] S. L. Vladimir Uversky, ed., *Flexible Viruses: Structural Disorder in Viral Proteins*. Wiley, 2011.

[264] S. J. Metallo, "Intrinsically disordered proteins are potential drug targets.," *Curr Opin Chem Biol*, vol. 14, pp. 481–488, Aug 2010.

[265] A. V. Follis, D. I. Hammoudeh, H. Wang, E. V. Prochownik, and S. J. Metallo, "Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules.," *Chem Biol*, vol. 15, pp. 1149–1155, Nov 2008.

[266] P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. L. Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp, and J. E. Bradner, "Selective inhibition of BET bromodomains.," *Nature*, vol. 468, pp. 1067–1073, Dec 2010.

[267] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker, *Natively Disordered Proteins*, pp. 275–357. Wiley-VCH Verlag GmbH, 2008.

[268] P. E. Wright and H. J. Dyson, "Linking folding and binding.," *Curr Opin Struct Biol*, vol. 19, pp. 31–38, Feb 2009.

[269] D. Hanahan, "Studies on transformation of *Escherichia coli* with plasmids," *Journal of Molecular Biology*, vol. 166, no. 4, pp. 557 – 580, 1983.

[270] D. C. Chen, B. C. Yang, and T. T. Kuo, "One-step transformation of yeast in stationary phase.," *Curr Genet*, vol. 21, pp. 83–84, Jan 1992.

[271] J. F. B. Leonard G. Davis, Mark D. Dibner, ed., *Basic Methods in Molecular Biology*. Elsevier, 1986.

[272] S. A. F. Sean Gallagher, Scott E. Winston and J. G. Hurrell, *Current Protocols in Molecular Biology*, ch. 10.8.1-10.8.24. John Wiley and Sons, Inc., 2004.