Selection Pressure on the Hepatitis B Virus in a Chronically Infected New Zealand Tongan Population

Brook Geoffrey Warner, BSc

THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
School of Biological Sciences

A Thesis submitted in fulfilment of the requirements for the Degree of Master of Science in Biological Sciences, The University of Auckland

February 2010

Abstract:

Introduction. In chronic hepatitis B virus (HBV) infections, the immune response at the time of HBeAg seroconversion usually suppresses viral replication, resulting in an inactive, HBeAg-negative chronic HBV infection (e-InD). However in some patients this anti-viral response does not suppress replication, resulting in an immune attack against liver cells known as HBeAg-negative chronic hepatitis B (e-CHB). e-CHB causes liver cirrhosis and liver cancer. The failure to suppress replication is probably due to a deficit in the immune response to the virus. We have used phylogenetic analysis of cloned HBV core gene sequences to determine whether there is a deficit in positive selection pressure on the core gene in e-CHB. This might give clues to the nature of any immune deficit in e-CHB that might be the target of a therapeutic vaccine.

Methods. We sequenced 196 clones containing the HBV core gene from 6 subjects with e-CHB and 10 HLA class I-matched subjects with e-InD. Measurements of genetic diversity and positive selection pressure made using PAML, PAUP*, MacClade, Geneious, and FigTree were compared in e-CHB and e-InD.

Results. There was a trend towards ($p=0.06$) greater intra-patient diversity in HBV sequences in e-InD than e-CHB. A PAML analysis detected six amino acid sites (21S, 26S, 77E, 113E, 130P, 180E) under positive selection pressure ($\omega=7.74$) within e-InD that were not found in e-CHB. An investigation of the number of subjects who demonstrated a non-synonymous change at these six amino acids revealed an almost significant p-value of 0.06. We were unable to elucidate any significant difference in selection on external branches, however we found a significant difference between internal branch lengths ($p=0.03$). Furthermore, we found evidence of HBeAg-mediated frequency-dependent selection in chronic HBV infection. This offers a novel perspective on the evolutionary mechanism of viral persistence.

Summary. We were unable to identify a large deficit in positive selection pressure on the HBV core gene in e-CHB relative to e-InD.

Conclusions. Our study serves as a pilot for a larger study, providing valuable information regarding power calculations, sample sizes, and case control matching regimes. Our data are consistent with a subtle immune deficit in active disease patients, and thus provide impetus for further investigation.

I wish to express my special thanks to:

The Auckland District Health Board Charitable Trust, for funding this research.

Dr. Bill Abbott (PhD) and Prof. Allen Rodrigo (PhD, DSc) - for their excellent supervision and instruction

Peter Tsai and Vicky Fan (Bioinformatics research programmers) - for their assistance with the many computer programs, and statistical analyses.

My family, for their encouragement, support and understanding.
My wife Renda, for her amazing moral support during this pivotal endeavour.

Table of Contents

List of Figures

List of Tables

ABBREVIATIONS:

HBV	Hepatitis B Virus
CHB	Chronic Hepatitis B
e-CHB	E antigen negative Chronic Hepatitis B
e-InD	E antigen negative Inactive Disease
CTL	Cytotoxic T Lymphocyte
HTL	Helper T Lymphocyte
NK	Natural Killer
IFN	Interferon
ORF	Open Reading Frame
HBp/e/s/x/cAg	Hepatitis B Polymerase/ E / Surface / X / Core Antigen
Ag	Antigen
HLA	Human Leukocyte Antigen
ALT	Alanine AminoTransferase
PC2	Physical Containment level 2
PEG	PolyEthylene Glycol
LB	Luria Broth
SOC	Super Optimal broth with Catabolite repression
PCR	Polymerase Chain Reaction
FDS	Frequency Dependent Selection
PAML	Phylogenetic Analysis of Maximum Likelihood
ML	Maximum Likelihood
NJ	Neighbour-Joining
GTR	General Time Reversible
HKY	Hasegawa, Kishino and Yano
SH	Shimodaira and Hasegawa
PAUP	Phylogenetic Analysis Using Parsimony
PhyML	Phylogenetic analysis of Maximum Likelihood
BLAST	Basic Alignment Search Tool
dN/dS	Δ Non-synonymous / Δ synonymous

Chapter 1:

Introduction

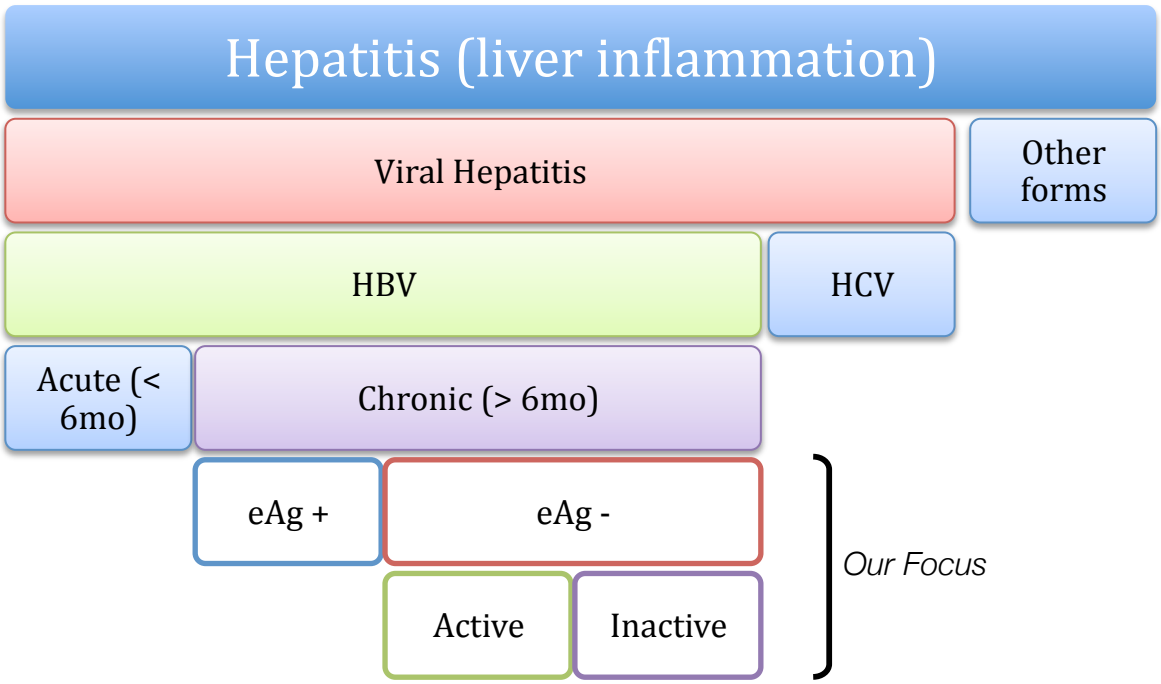
1 Introduction

1.1 Introduction:

It first is necessary to create distinction and clarity in the nomenclature regarding hepatitis infection, due to several variations being used interchangeably. Firstly, 'Hepatitis' refers specifically to liver inflammation, and is attributable to several agents, such as alcohol abuse as well as viral infections. 'Viral Hepatitis' is predominantly caused by Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV). It is important to note that HBV and HCV do not belong to the same viral family, and are named according to the target of infection (hepatocytes) [1].

Viral Hepatitis has two forms, acute and chronic, with the chronic form being further divided into an E antigen (eAg) positive and negative classification. Seroconversion from the eAg positive to the eAg negative form is caused by immune activity.

Within the eAg negative cohort, two further classifications exist – active and inactive disease. These two groups form the focus for our study, and are used as case and control, respectively.



It is accepted in the scientific community that 'chronic hepatitis B' refers specifically to the disease, i.e. – ongoing *hepatic inflammation*, where as 'chronic hepatitis B virus infection' refers to patients who exhibit ongoing *viral infection*. Although a subtle difference, it can cause confusion. Thus "Chronic Hepatitis B" describes chronic active disease, whilst "Chronic Hepatitis B infection" describes chronic inactive disease.

1.2 Chronic Infection

As mentioned, there are two classes of hepatitis infection:

- Acute Hepatitis B infection – is resolved by spontaneous viral clearance by the immune system, with the development of T and B cell sets that are specific for HBV antigens. This is observed in 95% adults, and in 10-30% of children [2]. Typically, adults acutely infected with Hepatitis B Virus (HBV) will develop a strong immune response to the virus and therefore clear the infection spontaneously. Our study does not focus on acute infections.

- Chronic Hepatitis B Viral Infection – this is presumed to be the result of an inadequate $CD8^+$ T-Cell response to viral exposure, and therefore leads to viral persistence in the host.

Patients exhibiting Chronic Hepatitis B Virus Infection fall into two classes – Active and Inactive disease. Inactive disease patients do not usually require treatment, since the virus is 'controlled' by the immune system. Active disease patients require treatment, and respond poorly to current treatments [2].

Chronic infection, or viral persistence, is a poorly understood phenomenon. Few risk factors that predispose individuals to chronicity have been characterised, and thus further research is required. A fuller understanding of the mechanisms allowing viral chronicity, and of the host–virus interactions, will bring us closer to developing an effective treatment for chronically infected patients.

1.3 Viral dynamics

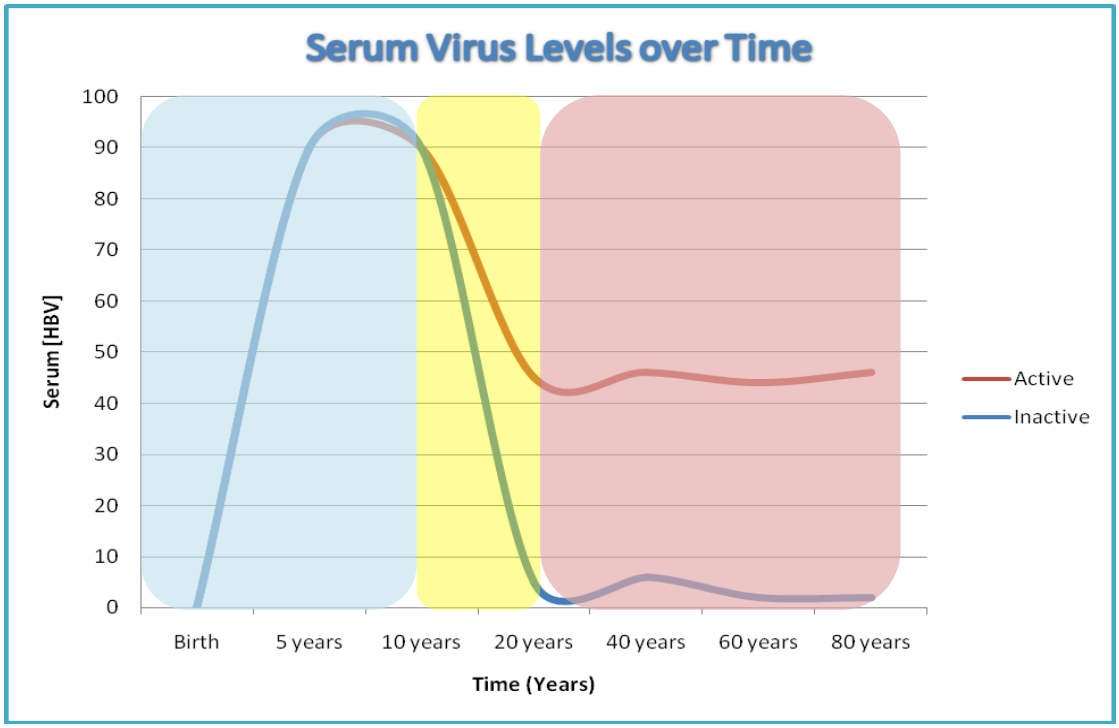
In a "typical" chronic (>6 months) HBV infection, the disease progresses through a series of three phases [3], which is thought to be facilitated by $CD4^+$ and $CD8^+$ T-cells:

- The inactive, eAg positive, 'immunotolerant phase'
- The active, seroconversion phase (loss of the e Antigen)
- The eAg negative, inactive disease phase (also referred to as e-InD) (See Figure 1:1, and see "Immunology" below).

However, in a large subset of patients, disease progression does not occur as outlined above. Following eAg seroconversion, viral replication is not adequately suppressed, leading to persistent immune attack on the infected hepatic cells and the resultant inflammation. This is classified as eAg negative, active disease (also referred to as e-CHB [e negative chronic hepatitis b]).

eAg seroconversion occurs for unknown reasons, and is often seen as a premature stop codon arising in the e-antigen coding region, usually at nucleotide 1896[4-6]. Until recently, this was assumed to be an indicator of immune control[7-9]. However, patients manifesting active disease are characterised by an E antigen negative viral genotype (HBeAg-), and yet show a high viral load in their serum indicating continuation of viral replication [10]. These patients have a particularly high progression to cirrhosis and therefore hepatocellular cancer, and respond poorly to current treatments [10]. $HBeAg^-$ Active disease possibly represents the largest subset of patients worldwide [2] and current treatments are problematic due to the high cost, the development of resistance, and the formidability of needing lifelong medication [10].

Figure 1:1 - Serum HBV levels seen as a function of time. Blue shading – Pre-seroconvertant, eAg positive, immunotolerant (no disease); Yellow shading – Seroconversion phase (loss of eAg); Red shading – Chronic phase, eAg negative.
Red line – High viral load, active disease; Blue line – Low viral load, inactive disease.



1.4 Aetiology, epidemiology and treatment:

The Hepatitis B Virus (HBV) causes all forms of Hepatitis B. It has been shown that the host immune response and subsequent influx of inflammatory mediators is the mechanism of damage, with the net result being damage to the hepatocytes [2, 11]. This is predominantly mediated by Cytotoxic T-Lymphocytes (CTL's), who initiate apoptosis in infected cells. The disease often progresses from inflammation to necrosis, and in 10-30% eventually to fibrosis and cirrhosis (irreversible damage) [11]. This highly toxic environment predisposes cells to genomic alterations, predisposing the individual to hepatic cancer [12].

It is estimated that 350 million people worldwide exhibit some form of persistent HBV infection, although this number is hard to calculate due to occult (silent) carriers [2, 11]. A third of these carriers live in the West Pacific region [13]. This encapsulates 5% of adults and between 40-90% of children and neonates [14] who do not adequately clear the virus, and thus manifest chronic infection [2, 15].

eAg negative, active disease (also referred to as e-CHB) is a major health burden, with 10-15% of the above population (chronic carriers) affected [11]. e-CHB is a leading cause of liver cirrhosis, with 25% progressing to terminal HCC [2, 10-12, 15]. 60% of e-CHB patients will develop cirrhosis within two years[6]. The risk factors for CHB are not as well studied but include: infection in the first two years of life [10], male gender, HBV genotype C, Diabetes, polymorphisms in the IFN-γ receptor, and immunocompromisation [6]. e-CHB is aggressive, and has a poor prognosis in response to treatment [[10]. The number of cancer deaths caused by HBV is thought to be approximately 320,000 per year, worldwide [10]. The disease continues to be a major health burden, despite a prophylactic vaccine being available [15]. This vaccine however is not 100% effective, and the high cost can prevent uptake in developing countries [16].

Currently there are several forms of treatment available. For chronic infections in the immunocompetent, anti-viral pharmaceuticals such as Lamivudine ® or Adefovir® (Nucleoside / nucleotide inhibitors) can be effective [10]. However, this form of therapy is not highly successful for carriers whom already have a weakened immune system, which is a possible cause of the initiation of chronicity [2, 11, 15]. To complicate matters further, resistance to these treatments has been shown to develop as rapidly as 12 months, which precludes their lifelong use [10]. These treatments are also only virostatic, rather than virocidal [14]. Patients with the HBeAg- Active disease phenotype (e-CHB) respond poorly to these treatments [2, 10, 11]. Alpha Interferon (IFN-α) is another available treatment. It is thought that IFN-α acts to boost the immune system, and promote the degradation of viral mRNA [6, 17]. However, due to the low immune activity of chronic HBV carriers, this treatment is also relatively ineffective [2, 10, 11]. A third and more dramatic option for late disease is liver transplantation. However due to the ongoing costs and complications, as well as the inherent lack of liver donations, this option is costly and impractical in endemic areas [1, 6].

1.5 Virology:

HBV is a non-cytolytic, hepatotropic virus, belonging to the family Hepadnaviridae (Hepa – Liver, dna – DNA)[2, 11]. HBV is the smallest DNA virus, and has a unique and complex genome [2, 18]. There are four known serotypes (adr, adw, ayr, ayw), which are based on antigenic epitopes presented on the envelope proteins[19]. There are eight genotypes (A-H) according to overall nucleotide sequence variation of the genome[19]. Our study focuses on genotype C and D.

Hepadnaviridae are a viral family that replicate via an RNA 'replicative-intermediate', and whilst this usually is a highly immunogenic molecule, the replication occurs within the nucleocapsid, preventing immune recognition [2]. This RNA intermediate also introduces a high mutation rate, which therefore allows for immune escape[20]. Typical DNA viruses does not demonstrate high mutation rates, explaining the ability of the immune system to control any novel 'variants' of such viruses (smallpox, varicella-zoster etc); and simultaneously explaining the inability of the immune system to control HBV. A study by Lok et al in 1996[5] demonstrated that the mutation rate in HBV increases as the individual progresses through seroconversion (See Table 1:1)

Table 1:1 - Mutation rate in HBV during the various phases.

	Pre-seroconversion (eAg+)	During seroconversion	Post-seroconversion (eAg-)
Nucleotide position/ year	0.4 ± 0.1	1.9 ± 0.3	2.4 ± 0.4
Codon / year	0.04 ± 0.02	0.21 ± 0.05	0.38 ± 0.07

Source: Lok et al - High rate of mutations in the hepatitis B core gene during the immune clearance phase of chronic hepatitis B virus infection; 1996; Hepatology, 24 (1), pp. 32-37

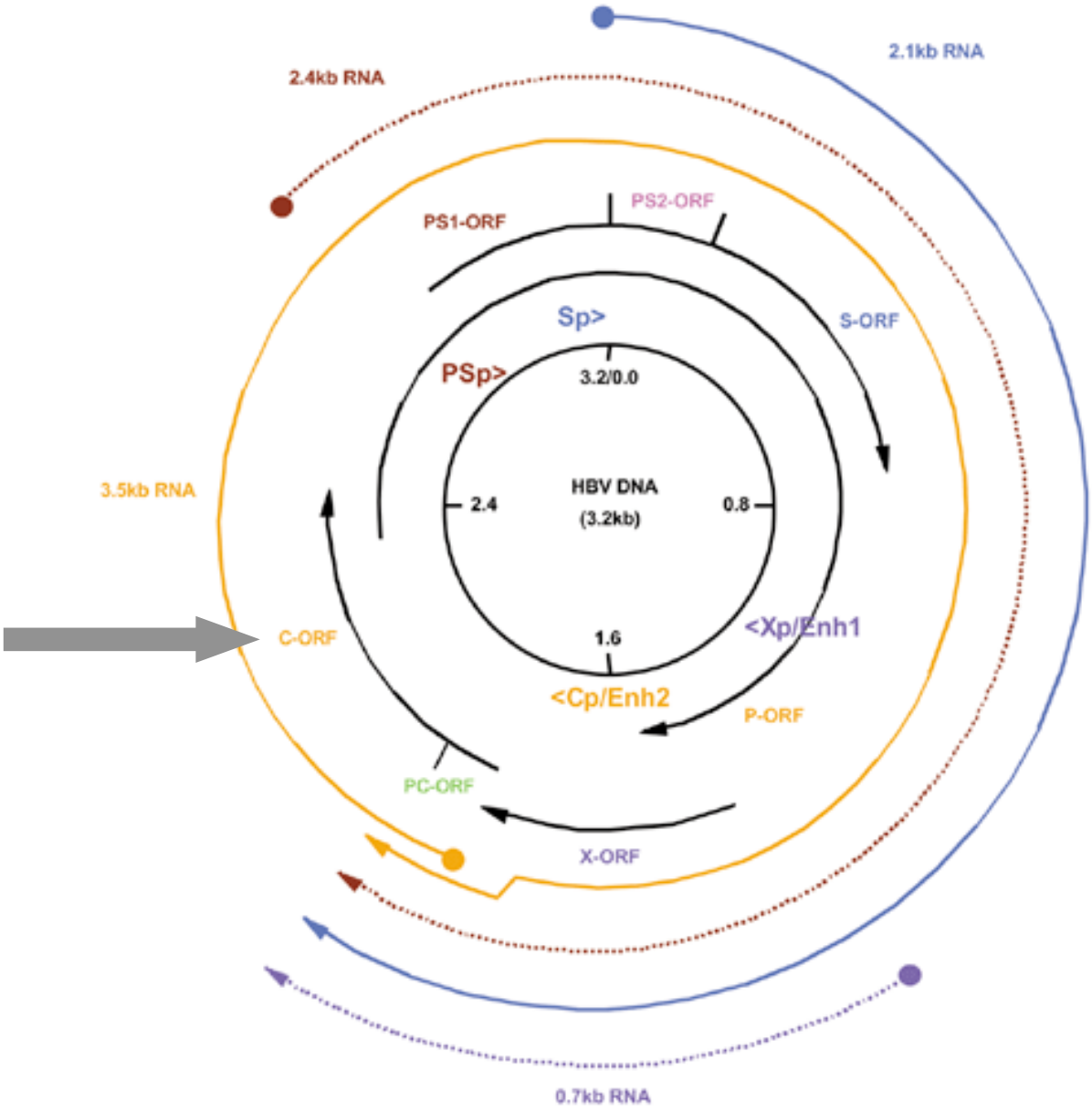


Figure 1:2 - HBV genome, showing overlapping reading frames.
The grey arrow highlights the Core ORF

"REGULATION OF HEPATITIS B VIRUS (HBV) GENE EXPRESSION" taken from:
Laboratory of Alan McLachlan, Department of Cell Biology, The Scripps Research Institute

Upon infection, HBV does not replicate efficiently immediately, but instead has a lag-phase of approximately 4-7 weeks [18, 21]. Whilst in serum, the minus strand is 3.2kb in length; the positive strand varies but is usually around 2kb [2, 11]. This creates a partially double-stranded genome, in what is known as a 'relaxed circle' (rcDNA) conformation [2, 11] (See Figure 1:2). Upon exit from the nucleocapsid, once inside the hepatocyte, this rcDNA is converted to cccDNA (covalently closed circle DNA [2]. Due to the relatively small size of the viral genome, Open Reading Frames (ORFs) are used to their full efficiency, with some degree of overlap [2]. The proteins encoded by HBV are polymerase (P), core (C), envelope/surface (S); as well as two proteins with incompletely understood functions – the X protein and the E antigen [11, 18]. It is thought that the X antigen plays an important role in hepatocellular carcinoma (HCC) by acting as a transcriptional activator of many cellular genes, including c-myc and p53 [2, 22]. The E antigen is a soluble protein generated by post-translational modification of the nucleocapsid protein, which itself is transcribed from the core gene. The E antigen plays a pivotal role in chronic hepatitis research. The exact function of the eAg is still relatively uncertain, however mounting evidence indicates a role in immunomodulation / tolerance [18, 23]. In mice, it has been shown that E antigen suppresses the Th1 response, thus eliciting inhibition of the $CD8^+$ T-cell response [2]. Likewise, the E antigen 'dampens' the T-cell response to stimulatory antigens[4, 6]. It is thought that there are multiple factors that contribute to this 'tolerising effect', such as T-reg cells, dendritic cell impairment, and the liver environment [18]. However, this putated immunosuppressive effect plays an important role in the viral evolution and replication dynamics. Frequency Dependent Selection (FDS), also known as Rare Allele Advantage, significantly modifies the impact of selection pressures on the virus population. This is covered in more detail below in section 1.7 - "Evolution".

1.6 The Immune Response:

Inactive disease

When patients effectively suppress the virus (Inactive Disease (e-InD)), the immune response is standard (See Figure 1:1) and follows a series of 3 phases. During the first 'immunotolerant' phase, serum HBV levels are high [3]. Lymphocyte numbers

and viral load are inversely proportional in this period.

In the seroconversion phase, the immune system actively recognises and elicits a response on the virus, and virus containing hepatocytes [3]. This manifests as a temporal period of sickness, usually with high ALT (alanine amino-transferase, a liver enzyme) levels. HBeAg and HBV DNA levels significantly decline, with an inversely proportional rise in anti-HBe antibodies [7]. In most cases, the loss of the E antigen is due to the appearance of a translation stop codon at site 1896 in the Core ORF[7]. $CD4^+$ T-cells are involved in the production of cytokines that stimulate $CD8^+$ T-cells, which are then employed to suppress viral replication via secreting IFN-γ [24]. B cells are also induced to produce antibodies that neutralise free viral particles, maintaining low viral titres. $CD8^+$ T-cell deletion experiments have indicated that T-cells are the predominant subset responsible for viral suppression [[25]. The activity of natural killer (NK) cells remains uncertain.

The third phase – 'inactive disease' is characterised by immune-mediated suppression of viral replication, and the absence of the E antigen. 'Flares' of disease in these patients can be fatal, but are uncommon [11]. Occasionally, some of these patients (1% per year) appear to lose the virus completely (HBsAg seroconversion). This is a little understood phenomenon. [7].

Active disease

In stark contrast, CHB has a different response pattern. At the time of seroconversion, the immune response seems markedly deficient, leading to only partial suppression of viral replication[6]. Interestingly, most patients will lose the E antigen (via the aforementioned appearance of a stop codon); yet maintain relatively high viral titres (>200,000 copies /ml) [[1]. In addition they exhibit raised ALT levels (>60U/L), and associated supportive liver changes[6]. Thus, the viral infection is classified as 'able to elicit disease', due to the high viral levels inducing an inflammatory response [7]. This group was the focus of our study - specifically of interest to us was whether there was a difference in selection pressure between this active disease cohort and the inactive disease cohort.

One of our aims was to elucidate whether the active disease cohort exhibited immune escape mutants in their serum, and then to characterise any mutations found. This would allow us to make inferences about the $CD8^+$ T-cell function.

Some attractive hypotheses have been proposed. It has been shown in previous studies [9, 14, 25, 26] that liver biopsies stained for T-lymphocyte presence show evidence of a large infiltrate, however it has also been shown that $CD8^+$ T-cells taken from the liver do not respond to HBV antigens *in vitro* [24]. It has been suggested that perhaps the dysfunction may be in a CD4 – CD8 feedback mechanism, thus leading to unregulated stimulation of $CD4^+$ T-cells[6]. This has lead to the putative hypothesis implicating the $CD4^+$ T-cells in eliciting the damage observed in hepatitis liver inflammation [24].

It is also thought that in order for the virus to persist in the body, there are mechanisms allowing the co-existence of functionally active $CD8^+$ T-cells and high levels of viral antigen, without inducing an extreme response [27]. This may be due to viral strategies to avoid immune recognition, or the exhaustion/apathy of antigen specific $CD8^+$ T-cells [18, 27].

The 'tolerising effect' potentially also has a role in chronicity, although this is uncertain [18]. Viral tolerance could theoretically occur when an individual is infected during thymic education (i.e. - the first two years of life), leading to a state of tolerance towards HBV antigens [18]. The immuno-suppressive effects of the e antigen also need to be more fully examined, since this modifies the selection pressures acting on the viral genome.

While these mechanisms do exist, it is thought that this co-existence of active $CD8^+$ T-cells with viral antigen leads to selective pressure being placed upon the virus, creating immune escape variants [14]. It is presumed that the main mechanism allowing the continuation of viral replication, and thus infection, in active disease patients is some form of $CD8^+$ T-cell hypo-responsiveness or dysfunction.

Interestingly, Nowak et al proposed a mathematical model in 1996[28] which demonstrated the role of equilibrium and viral replication and recognition. This model showed that viral persistence could be a function of variations in the individual's CTL *responsiveness*, whilst still maintaining the same levels of overall CTL *response* [15, 27, 29]. Studies have demonstrated that patients with 'self limited Hepatitis' (inactive disease) have an increased $CD8^+$ response than patients

manifesting chronic disease [11]. Patients with active chronic disease, show reduced $CD8^+$ numbers in their serum [11].

1.7 Evolution:

To understand the evolution of HBV within the Tongan population, it is helpful to discuss what pressures are acting on the virus.
When an infection is at high viral titre, selection acts mainly on replication efficiency, due to the effective 'competition' between virus quasi-species for an uninfected cell. Therefore, any virus that can replicate more efficiently will gain advantage over less efficient viruses. Conversely, when the infection reaches low viral titre, selection will act on the viruses' ability to evade the immune system. Any virus containing a mutation that evades immune detection will be favoured and increase in number. Thus in a low titre environment, viral mutation is beneficial, and detrimental in a high titre environment[28]. It has previously been shown that in patients with e-CHB, amino acid changes within region 18-27 of the Core gene are able to inhibit the activation of Core specific $CD8^+$ T-cells [11, 26, 29], and thus may allow evasion of the immune system. This study aimed to expand on this result by elucidating further amino acids that enable immune evasion.

Theory of equilibrium:

As mentioned earlier, the putated immunosuppressive effect of the eAg needs to be examined. If the assumption that the eAg elicits immune suppression is correct, this introduces the possibility of an 'equilibrium' existing within a host.

Theoretically, during the infection, the virus will move through three stages. During stage 1 (at the beginning of the infection) the virus population contains the eAg and thus suppresses the immune response. During phase 2, (seroconversion) mutations begin to appear in the eAg of some viral clones, thus making it non-viable, abolishing its function and therefore allowing immune recognition. Conceivably, during this phase, two cohorts of virus now exist in the population - those who have acquired a mutation in the eAg, who are now known as eAg negative; and those who still retain eAg functionality (eAg+). Theoretically, this allows the eAg negative clones to 'cheat', by relying on the immunosuppressive

effect of the eAg positive clones[30].

By not producing eAg, these viral clones gain a replicative advantage, as every protein produced by the virus costs, metabolically. Simultaneously however, the virus risks placing itself at a survival disadvantage, as the eAg affords protection from the immune system. The notion of 'cheating' becomes of interest when a mixed population of eAg+ and eAg- exists within the patient. In this phase, eAg- viral clones *will also benefit* from their sister eAg+ clones that are still producing the eAg, and suppressing the immune response[30]. The question then becomes - what effect does this equilibrium have on selection pressure?

Nowak *et al's*[28] seminal paper, detailing a mathematical model relating to chronic viral infection dynamics, has been cited in the literature 357 times. Long *et al*[31] modified this model to apply to HBV in 2008. To the best of our knowledge there is no further model in the literature to date that shows the effect of frequency dependent selection on the viral population due to this immunosuppressant effect of the eAg.

What has not been well documented is the immune selection pressure acting on the virus population in CHB patients, focussed solely on e-InD versus e-CHB. Our study investigates these problems.

1.8 Focus and structure of this Thesis

This thesis seeks to elucidate differences in selection pressure between cases (active disease) and controls (inactive disease), and is structured into 6 chapters.

Chapter 1 covered the introduction to Hepatitis, as well as HBV, including aetiology, epidemiology, virology, genetics, evolution, and treatment. Chapter 2 covers the research rationale, study design, and the methods used at the PC2 lab at Auckland Hospital. Chapters 3 and 4 focus on the phylogenetic and statistical analyses performed at the Bioinformatics Institute. These chapters are structured into four sub-sections: introduction, materials and methods, results, discussion.

Chapter 3 demonstrates the use of the sequence alignments to remove clones containing nonsense mutations, details the resultant phylogram, and investigates the diversity found within the dataset.

In chapter 4 we examine the selection pressure acting on the whole tree, and then focus on the specific sites within patients that are under positive selection. We then conduct separate analyses on between-patient and within-patient selection pressures.

Following this chapter 5 discusses a novel insight into viral dynamics pertaining to the E antigen, which we came across in the course of our study.

The study summary and conclusions are found in Chapter 6, followed by the Appendices.

Chapter 2:

Research Rationale & Study Design

2 Research Rationale & Study Design

HBeAg- Active disease (e-CHB) probably represents the largest subset of patients worldwide [1, 2], and this is a particularly problematic group. Compared to inactive disease patients (both eAg+ and eAg-), e-CHB patients have a particularly high progression to cirrhosis and cancer, respond poorly to treatment, and have far more aggressive disease [11, 12]. Current treatments are problematic due to the high cost, the development of resistance, and the formidability of needing lifelong medication. The need for a more effective treatment of CHB is obvious.

Immunotherapy research has progressed significantly in the last decade, and it is thought that a therapeutic vaccine may hold the key to effective treatment of CHB. If an inherent immune deficit exists in patients with e-CHB, then it is an attractive possibility that a therapeutic vaccine may induce proper immune function and rectify this deficit. This would not only treat the disease but would have significant implications for management of the disease, post-treatment. This would reduce the economic and social burden of the disease, as it would not require long term monitoring programs or lifelong medication to control infection and disease [1, 10].

2.1 Aim and Hypothesis

Our hypothesis was that "Patients with HBeAg- active disease will display less CD8+ T-cell function and pressure than in e-InD, resulting in fewer mutations." This hypothesis was tested by comparing the ratio of non-synonymous to synonymous mutations (ω), in cloned viral genomes from each of the two groups. We aimed to demonstrate that there was mutational evidence that was consistent with $CD8^+$ T-cell dysfunction.

It was thought that the patients experiencing chronic HBV infection have some form(s) of $CD8^+$ T-cell dysfunction, possibly in feedback mechanisms existing between $CD8^+$ cells and $CD4^+$ T-cells, in molecular structure, or in the ability to recognise known immune epitopes [24]. Previous studies have demonstrated a lack of reactivity to HBV peptides by $CD8^+$ T-cells isolated from the liver of patients with chronic disease [24]. Our proposed therapeutic value would lie in the proposed ability to induce proper function of these T-cells, thus converting *active disease* to

inactive disease. There is some recent research into therapeutic vaccines, with some preliminary success [14]. There have also been reports of active- to inactive-disease conversion following bone marrow transplants, indicating feasibility [32]. A therapeutic vaccine must aim to stimulate proper $CD8^+$ T-cell function whilst not over-stimulating the immune system, as this will increase the disease rather than aid it. We aim to build on the recent advances in therapeutic vaccine development.

2.2 Study design rationale

$CD8^+$ T-cells are notoriously difficult to study. Their behaviour is somewhat temperamental and specific cytokine and antigen concentrations are required to achieve stimulation[1]. Study is further complicated by the vast array of HLA haplotypes. Therefore, to facilitate any study upon $CD8^+$ T-cells, one must know the HLA status of the patient from whom the sample was taken. Moreover, $CD8^+$ T-cells respond to a very limited range of epitopes, thus finding the right one from ~300 within the HBV viral genome is an expensive and laborious process. This also means that it is near on impossible to study each epitope with much detail, leading to an inherent lack of sensitivity [1].

We therefore decided to take a different approach to this research, by focussing our study design around investigating the viral genetic evolution, instead of directly testing the host immune response. This was achieved by comparing the mutations found in viruses extracted from active disease patients to virus extracted from inactive disease patients. One benefit of this approach is that it elucidated immune escape mutants, implying that they had presumably responded to an existing, functional, active T-cell.

We hypothesised that there was an inherent failure in the $CD8^+$ T-cells in the patients with e-CHB, and that by examining and documenting the mutations that arise because of immune selection pressure in both cohorts, we could make inferences about $CD8^+$ T-Cell function. We extracted, cloned, sequenced, and then phylogenetically analysed the viral genome isolated from 16 patients. We then conducted an analysis looking for patterns that emerge which gave indications or evidence of trends within each subset.

2.3 Subjects and Methods:

Our study population were all of Tongan ancestry, as this conferred four advantages. Firstly, HBV is endemic in this ethnicity (99% prevalence) and most individuals are infected by age 5, either vertically or horizontally [33]. Within the Tongan population, 10% exhibit chronic HBV infection [34].

Secondly, due to low genetic diversity within HLA loci, it is possible to match Tongan individuals at a minimum of 5 out of 6 HLA class I loci. 75% of our study population have the HLA-A*2402 allele [1]. Potential reasons for the low genetic diversity are thought to be selection pressure introduced by early European visitors to the West Pacific, and a possible founder effect during early migration from Asia [35]. In addition, strong linkage disequilibrium means that individuals matched for HLA-B will generally also be matched for HLA-C [1, 35]. This allowed us to compare individuals based on class matching their HLA alleles. This was important as it is thought that different HLA class-I alleles will induce different mutation repertoires, and creates an invariant context to examine viral mutations.

Thirdly, the C3 HBV genotype, a known risk factor for chronicity is the predominant form in Tonga. Fourthly, fortuitously, a Tongan specific HBV clinic has been established in a nearby suburb, permitting the easy recruitment of Tongan patients.

Cohorts

In this manner, there were three main cohorts within our sample population (See Table 2:1).

1. Those with E antigen positive inactive disease, and no current immune response (labeled in grey, Table 2:1). This is the 'immunotolerant phase', indicating that their immune system has not yet recognised the virus as foreign. Therefore they have a high viral load (20x104 - 20x109 IU/mL), and are consequently highly infectious [11].

2. Those with E antigen negative active disease (e-CHB) (labeled in red, Table 2:1). It is thought that these patients exhibit immune dysfunction, as the viral replication is not adequately suppressed. Our aim was to investigate whether there existed sufficient pressure to create immune escape mutations. It was thought that this 'middle ground' represented the adaptive advantage for the virus, as a high number of amino acid changes often results in a dysfunctional protein. Thus, by existing in this middle zone, the virus both escapes immune pressure and continues to replicate. This group had a lower viral titre than cohort 1 ($2x10^3$ - $20x10^6$ IU/mL) [11].

3. The third cohort (labeled in blue, Table 2:1) was comprised of individuals who have E antigen negative inactive disease, (e-InD) and whose immune system had suppressed viral replication. This final group showed evidence of immune escape mutation. This group also showed levels of HBV antigen specific $CD8^+$ T-cells, which is absent in the second cohort [11]. Occasionally (4-20%), some of these patients revert to active disease status (cohort 2) [11].

By using computational methods to convert mutation repertoires into a statistical form, we were able to analyse whether a viral clone is under immune pressure, or whether the mutation is a result of genetic drift.

Table 2:1 - Cohorts found within HBV. *Grey* – Pre-seroconvertant patients, e antigen positive, this cohort forms our 'baseline'; *Red* – Post-seroconvertant, e antigen negative, with a high viral load and thus chronic disease, this cohort forms our 'case' population; *Blue* – Post-seroconvertant, e antigen negative, with a low viral load, and inactive disease, this cohort forms our 'control' population.

Chronic HBV Infection Patient Cohorts				
	eAg Positive	**Alternate Title**	**eAg Negative**	**Alternate Title**
Very High HBV load	Inactive ($CD8^-$, $CD4^-$)*	Immunotolerant	None Known	NA
High HBV load	Active/Inactive ($CD8^-$, $CD4^+$)	Seroconvertants	Active (CD8-, CD4+)*	e-CHB
Low HBV load	Data unclear‡	NA	Inactive (CD8+, CD4+/-)*	e-InD

2.4 Study Phases

Phase 1 - Cloning and sequencing of HBV

The aim of phase 1 was to extract, amplify, clone, and sequence HBV genotype C3 and D4 clones from Tongan patients. (See Methods below).

Phase II - Bioinformatic modelling (phylogenetic) of generated data

The aim of phase II was to take the generated sequence data and conduct an analysis related to selection pressure, mutation repertoires, and viral diversity. (See Chapters 3 & 4 below) The main focus was to look at the types of substitutions and ascertain what 'pattern' of mutation and selection is present, and thus determine if the T-cells are functioning in an efficient manner in suppressing viral replication.

2.5 Possible results:

In regards to mutation patterns, we anticipated seeing a pattern that was consistent with partial immune pressure, or incomplete suppression, similar to previous observations in the S ORF. This is best demonstrated as a phylogenetic tree. (See Figure 2:1, Figure 2:2, and Figure 2:3).

2.5.1 Pattern 1: (E+ = e-CHB) < e-InD

The simplest mutation pattern would be if the number of mutations (defined by branch length) in the e-CHB [Red] cohort was very low, and equal to the number of mutations in the HBeAg positive cohort (baseline). This would indicate that there was no active selection pressure on the case population, thus inferring that no CD8+ T-cells had been generated. The problem with this result is that this would not explain why E antigen seroconversion had taken place without CD8+ T-cell pressure. This result is unlikely due to the acquisition of a stop codon in the eAg during seroconversion indicates some degree of immune activity.

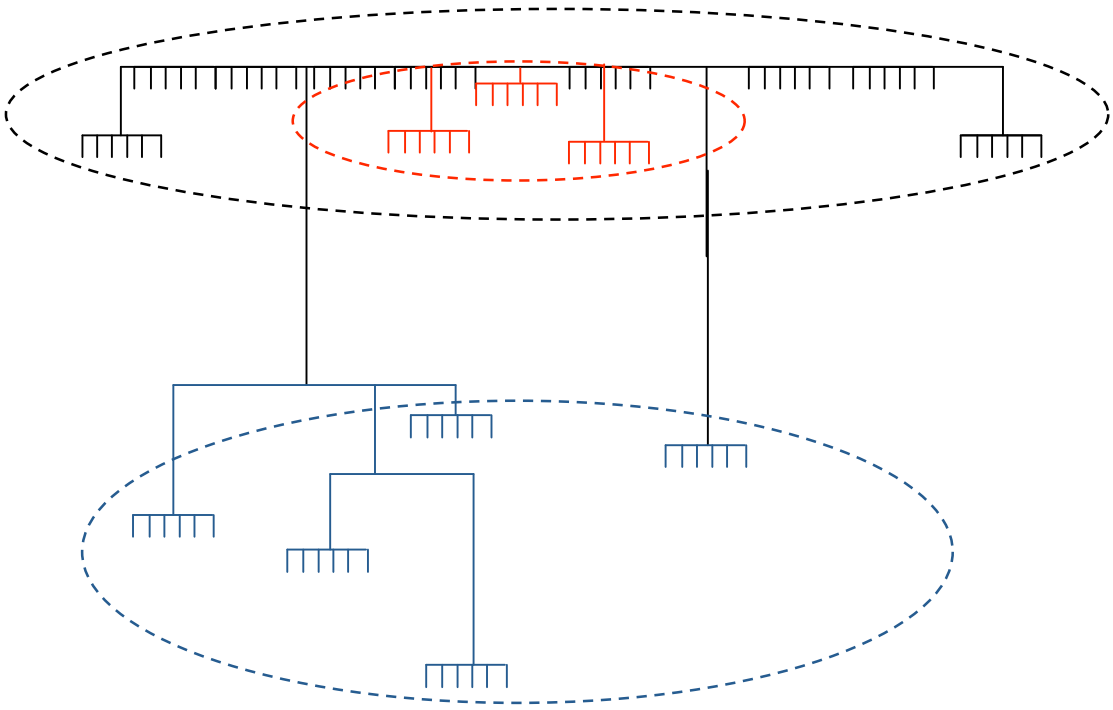


Figure 2:1 – Pattern 1 (eAg+ = eAg- CHB) < eAg- InD

Figures 2:2 – 2:4 – Sample phylogenetic trees showing potential patterns. The length of the vertical branches represents amino acid changes. The black circle encloses eAg+ patients (pre-seroconvertants). The red circle indicated eAg- CHB patients. The blue circle encloses eAg- InD patients.

2.5.2 Pattern 2: e+ < e-CHB < e-InD

A second possibility was that the number of mutations would be greater in the HBeAg negative, active disease cohort (e-CHB - [Red]) than in the HBeAg positive, inactive disease cohort [Black], and less than in the E-InD cohort [Blue] This would indicate that the CD8+ T-cells in these patients were eliciting some selection pressure upon the virus, and therefore induced some escape mutations, albeit less than in the E-InD cohort. This would indicate incomplete suppression - either of the virus by the immune system, or of the immune system by the eAg.

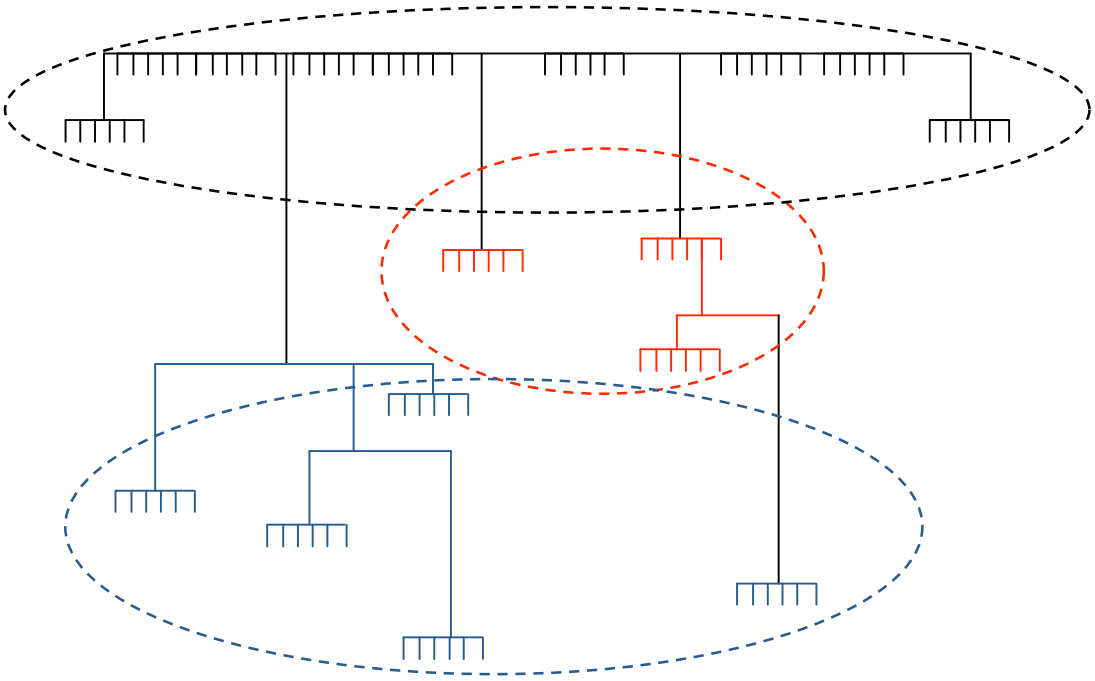


Figure 2:2 - Pattern 2: eAg+ < eAg- CHB < eAg- InD

2.5.3 Pattern 3: e+ < e-InD < e-CHB

A third possibility was that the e-CHB cohort demonstrated longer branch lengths than the e-InD cohort. This would have indicated that there was a high level of immune pressure in the e-CHB patient acting on the virus, thus causing the virus to mutate frequently.

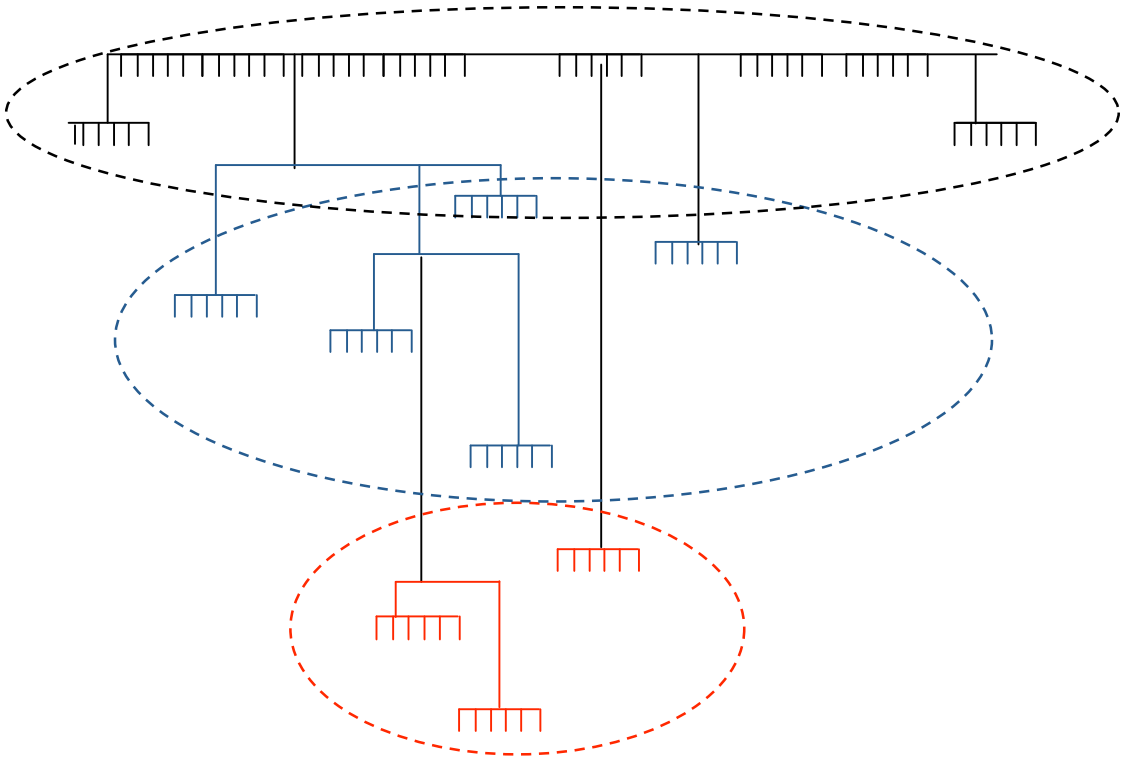


Figure 2:3 - Pattern 3: eAg+ < eAg- InD < eAg- CHB

Statistical methods would then be employed to analyse whether these mutations were a result of genetic drift or other stochastic processes, or were in-fact due to selection pressure. One way this could be achieved is by comparing the number of non-synonymous to synonymous changes. Findings would need to be confirmed by further $CD8^+$ T-cell studies.

For example, a high prevalence of positively selected substitutions would infer that the T-cells in the patient were acting upon the virus, yet the viruses were able to persist. This observation in a patient with active disease would also infer immune escape. This could be caused by the region of recognition by the T-cell Receptor (TcR) not eliciting 100% viral suppression, or alternatively mutation of immune epitopes may alter the immune response. A third possibility is immunodominance, a phenomenon that is observed when viruses are able to persist within a host. It refers to the tendency of the immune response to focus narrowly on one epitope, and ignore the rest. Often, this focus will change, thus allowing the virus to persist [29, [36]. The prevalence of this in CHB falls outside the scope of this study, however should be investigated further.

The possibility always exist that the immune response is specific for another HBV protein (not the core) and therefore this requires further investigation into other possible candidate proteins.

The fourth possibility was that no pattern emerged, which would indicate that perhaps CD8+ T-cell pressure was not responsible for the clinical differences between active and inactive disease. Therefore the working hypothesis is excluded requiring us to develop a new hypothesis.

2.6 Methods

This research was conducted both at the Liver Transplant Unit Research Group, on level 14 of the Auckland City Hospital Support Building, and at the Bioinformatics Institute, School of Biological Sciences, University of Auckland. All statistical software was either freeware, or was available at the Bioinformatics Institute, courtesy of Professor Allen Rodrigo. Ethics approval for this project had previously been granted from the New Zealand Ministry of Health (NTX/05/12/160)

2.6.1 Patient Selection and Recruitment:

Each year, the New Zealand Hepatitis B Screening Program regularly screens known HBV carriers to investigate if treatment is due or liver transplantation is required [1]. From this list, we isolated 345 Tongan carriers. Upon consent, blood samples were taken, (see below) allowing genotyping of the patient's HLA-A, HLA-B and HLA-C loci, and extraction of virus DNA from serum (see below). Of these 345, we isolated 8 who were E antigen negative, yet exhibiting chronic Hepatitis B (i.e. – e-CHB). This is defined as two elevated ALT levels, 3 months apart, detectable HBV DNA levels, and a positive Liver Biopsy. This subset is hereafter referred to as the 'case'. Two subjects were excluded due to having a rare HLA haplotype. A second control subset of patients was also selected. These patients exhibited e Antigen Negative Inactive Disease (i.e. - e-InD), defined as having a normal ALT level over an extended period. This subset is hereafter referred to as the 'control'. This allowed us to create a comparison regime between the two groups. Case patients (with active disease) are matched against at least two control patients (inactive disease) and at a minimum of 5 out of 6 HLA molecules. This allows the comparison of whether there are different mutation repertoires induced in the different subsets. To the best of our knowledge the patients are not related.

Table 2:2 - Clinical data from each subject, demonstrating their classification criteria

	Case							Control 1								Control 2							
	ID	Age	Sex	ALT level 1	ALT level 2	HBV DNA level	Liver Biopsy	ID	Age	Sex	ALT level 1	ALT level 2	ALT level 3	ALT level 4	Years	ID	Age	Sex	ALT level 1	ALT level 2	ALT level 3	ALT level 4	Years
1	16	62	M	154	84	>2e5	G1S2	365	39	F	16	15	22	8	7yrs	553	40	M	26	23	27	33	8yrs
2	29	58	F	257	319	>2e5	G2S2	305	36	F	7	33	29	26	4yrs	459	56	F	35	27	30	26	9yrs
3	249	44	M	66	113	>2e5	G4S4	290	46	M	54	48	39	46	7yrs	368	35	F	22	25	18	19	8yrs
4	250	39	M	81	78	>2e5	G3S3	8	42	M	31	27	25	27	7yrs								
5	308	51	M	241	63	24000	G3S4	413	48	F	33	25	24	36	8yrs								
6	318	66	F	173	311	>2e5	G2S3-4	337	61	F	41	29	32	30	7yrs								

NB - Control 3 for group 1: Patient 569 - Age: 47, Sex: Female, ALT1: 37, ALT2: 19, ALT3: 22, ALT4: 24, Years: 7

Table 2:3 - HLA haplotype of each subject

		Case				Control 1				Control 2		
		HLA-A	**HLA-B**	**HLA-C**		**HLA-A**	**HLA-B**	**HLA-C**		**HLA-A**	**HLA-B**	**HLA-C**
1	**016**	1101/2402X	4001/5602	0102/0304	**365**	0206/2402	4001/5602	0102/0304	**553**	1101/2402	4001/5602	0102/0401
2	**029**	0212/2402	5502/5602	0102/0102	**305**	0206/2402	5502/5602	0102/0102	**459**	0206/2402	5502/5602	0102/0102
3	**249**	2402/2402	4801/5602	0102/0801	**290**	2402/2402	4801/5602	0102/0801	**368**	2402/2402	4801/5602	0102/0801
4	**250**	2402/2402	3901/5602	0102/0702	**008**	2402/2402	3901/5602X	0102/0702				
5	**308**	1101/2402	1506/4010	0403/0403	**413**	0206/1101	1506/4010X	0403/0403				
6	**318**	1101/3401	4002/4801	0801/1502	**337**	1101/3401	4002/4801	0801/1502				

NB - Control for group 1: Patient 569 - HLA-A: 2402/2601, HLA-B: 4001/5602, HLA-C: 0102/0304

2.6.2 Serum Extraction

Extraction of Blood from patients

Blood samples were collected by registered nurses who collected 3mL of whole blood for DNA analysis, and 2 mL of sera for viral extraction.

Blood for DNA analysis was collected into heparin tubes, containing 10mL of anti-coagulant. 40mL of sucrose lysis buffer was then added to induce erythrocyte swelling and lysis via osmosis. Tubes were then spun at 2,000x g for 10 minutes, and the lysate was discarded. The remaining product is then used for HLA allelotyping (See HLA Allelotyping below)

Blood taken for serum samples was collected into serum collection tubes, which was then spun at 2,000xg for 10 minutes, and the resultant supernatant collected and stored at -80℃.

2.6.3 Viral Extraction

Isolation of Virus from patient samples:

HBV DNA was extracted from patient serum with the Roche High Pure Viral Nucleic Acid kit, according to the manufacturers instructions, with the following modification: The first spin was changed from one minute at 8,000xg to two minutes at 2,000xg, to ensure that all virus present in the sample has ample chance to bind to the filter.

Amplification of Virus DNA by Polymerase Chain Reaction

Extracted samples were stored at -20°C. Due to between- and within – patient differences, primers often required customisation to amplify e-CHB HBV DNA. Thus, we obtained a consensus sequence of base pairs 1700-2390 of each patient, which allowed us to design primers to fit. These were 5'-TTCACCTCTGCACGTCG, 5'-ATGTCAACGACCGACCTTGA and 5'-GAGGCTGTAGGCATAAATTGGTCT. All have matching reverse primers. To prevent PCR contamination, the products and reaction are kept apart. Setup is undertaken in a PCR hood, with the template added in separate room. The PCR products are handled and analysed in a third area.

High Fidelity Accuprime Taq was used to ensure accurate amplification for phylogenetic analysis. PCR products were run by gel electrophoresis on an ethidium bromide agarose gel, and samples that yielded a strong band were then PEG cleaned (see method below) and re-suspended in 4.0µL of double distilled H_2O (dd H_2O). Samples which did not yield a band, including the negative control, were PEG cleaned, re-suspended in 3.0µL ddH_2O, and then PCR amplified using a second set of internal primers, 5'-TCACCTCTGCACGTCGCAT and 5'-CGCTGCGTGTAGTTTCTCTCTTATA

The primers are designed to amplify the core open reading frame, beginning at a highly conserved region in all HBV known as DR1 [bp=1901]. The 3' primer is located at a conserved sequence within the Tongan C3 genotype, about 100bp into the Polymerase ORF. The total length of this fragment is 1140bp. Therefore this fragment contains the distal X ORF, the core promoter, the entire C ORF and a few hundred base-pairs of the P ORF.

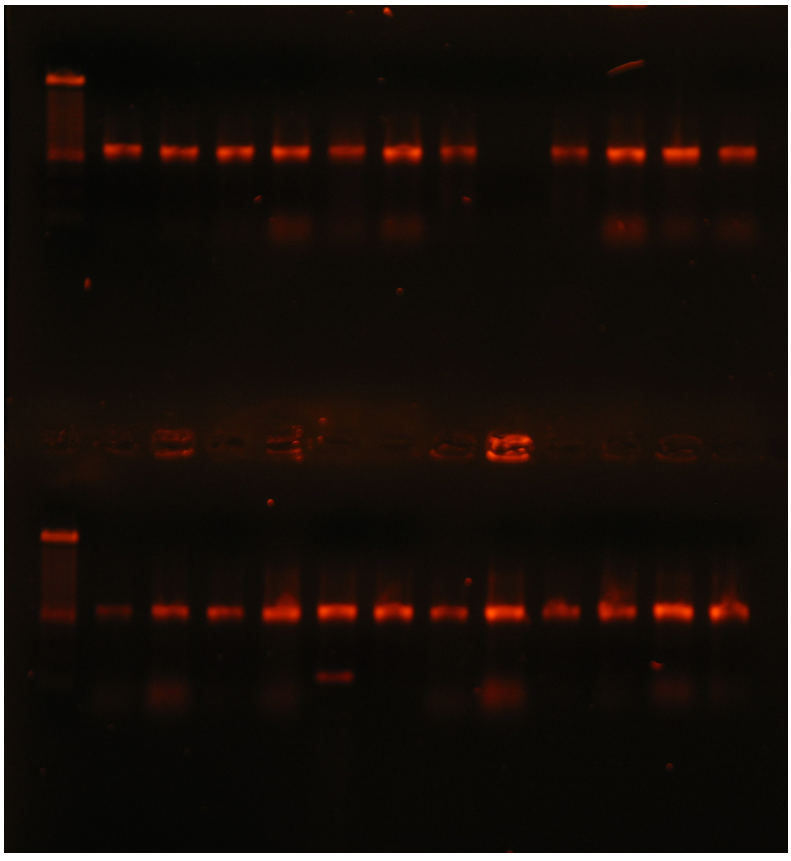


Figure 2:4 - Ultra-violet photo of PCR Electrophoretic gel, Colony Pick using the primers 5'-CGCTGCGTGTAGTTTCTCTCTTATA and 5'-TCACCTCTGCACGTCGCAT. Bold lanes were selected for sequencing.
Top row, left to right: Marker, **016D**, 016D, 016D, 016E, **016E**, 016E, 016E, Negative Control, **016F**, 016F, 413G, **413G**, 413G;
Bottom row, left to right: Marker, 413H, 413H, **413H**, 308C, 308C, **308C, 308D**, 308D, 308D, 308E, 308E, **308E**.

2.6.4 Cloning

A-tailing:

Following a successful PCR, the amplimer was PEG purified, and resuspended in 4.0µl ddH_2O. 4.8µl of A-tailing solution was added, and samples were heated to 72°C, and spun every 8 minutes to mix the resulting condensate. A-tailing ensures that the ends of the core ORF fragment are 'sticky' to allow efficient incorporation into the cloning vector.

Ligation:

Following the A-Tail, samples were again PEG purified, and resuspended in 4.8µl ddH_2O. Samples were then added to a pGEM-T ligation reaction, at a Vector : Insert ratio of between 1:1 and 1:10. Samples were left to ligate for >24 hours, at 4°C.

Production of Competent Cells:

220mL of Terrific Broth was prepared by adding 2.2212g Glycerol and 10.472g Terrific Broth Medium to 220mL ddH2O, then autoclaved. Escherichia Coli DH5α cells were streaked onto an LB plate (without Ampicillin) and grown overnight at 37°C. 5mL of Terrific Broth was aliquoted into two 50mL Falcon tubes, in a sterile Class II hood. A single colony from the overnight culture was picked and re-suspended in each aliquot. 5µl of 100mg/ml Ampicillin (100 µg/ml final concentration) was added to the first tube, which was labelled as a control. Both tubes were incubated at 37°C, 200 RPM overnight, with the Falcon tube lids slightly loosened to allow airflow. 1mL of Terrific Broth was set aside for use as a blank.

2mL from overnight culture was added to remaining Terrific Broth (~210mL), and incubated for 90mins, at 200RPM. Optical Density of 0.3 was ascertained using a spectrophotometer, blanked against Terrific Broth without E.Coli. Cells were placed on ice to arrest replication. 200mL culture was then split into four 50mL aliquots, and centrifuged at 3800RPM for 15mins at 4°C. The supernatant was discarded, and the bacterial cell pellet was resuspended in 50mL of ice-cold 50mM CaCl2 (12.5 mL per tube). Cells were placed at 4°C for 20 mins, aliquoted

into 1.5mL centrifuge tubes, and snap frozen using ethanol and dry-ice. Cells were stored at -80°C until required.

Transformation:

Transformations were conducted according to the following protocol: Agar plates were pre-warmed at 37°C for 30 mins, and then inoculated with a Water / Ampicillin Mixture, and placed in the incubator for 1 hour. At 1 hour the plates were removed and allowed to dry, with the lids slightly ajar. Concurrently, required volumes of competent cells (E.Coli DH5α) were thawed on ice. Once thawed, 50µL of competent cells were added to each ligation reaction (see Ligation above), and stood on ice for >30 mins. Cells were then heat-shocked at 42°C for 50 seconds, and then stood on ice for a further two minutes. 950µL of SOC media was then added to each reaction, and reactions were placed in the shaking incubator (150rpm, 37°C) for 90 minutes.

At 90 minutes, the cells were spun at 13,000xg for 10 seconds, and 800µL of supernatant was discarded. The cell pellet was then re-suspended by gentle pipetting in the remaining 200µL, which was then spread on the agar plates. Following an overnight incubation, transformed colonies were seen on the agar plates. Growth on the agar is mediated by the ampicillin resistance gene contained in the pGEM-T vector (See appendix). A single colony from each plate was selected and PCR amplified, using the primers 5'-CGCTGCGTGTAGTTTCTC TCTTATA and 5'-TCACCTCTGCACGTCGCAT.

1.0 µL of this amplimer was subsequently run on an ethidium bromide gel, and reactions that produced a band greater than 800bp were then selected for analysis. Samples were again PEG purified, and re-suspended in 8µL of ddH_2O. This sample was then Nanodropped to ascertain DNA concentration, ready for sequencing.

2.6.5 Sequencing

Sequencing of Virus ORF:

Sequencing of the viral clone took place at the University of Auckland, School of Biological Sciences, Centre for Genomic Processing. Samples were PCR amplified using only an upper or only a lower primer. The amplimer is then 'cleaned' for loading into the sequencing machine according to the manufacturers instructions[37].

2.6.6 HLA Allelotyping:

A system of elimination was used to ascertain the alleles that were present in each 'case' and 'control' patient studied, based on known polymorphisms in the HLA gene. Geneious©[38] was used to construct a custom BLAST database of known HLA-A, HLA-B, and HLA-C allelotypes. The raw sequences were then edited to include ambiguity codes for all sites of heterozygosity. These edited sequences were then BLASTed against the custom database, and the allelotypes with the lowest e-value and bit-score were noted. Following this, the edited sequences were then used to eliminate non-matches, according to the method described by W Abbott *et al*, 2006 [35].

Any ambiguities that could not be resolved using this method were then selected for the cloning process, where one of the two alleles is incorporated into pGEM-T and sequenced, thus allowing the resolution of the ambiguity

Chapter 3:

Alignments, Phylogenetics, and Diversity

3 Alignments, Phylogenetics, and Diversity

3.1 Introduction

Our original hypothesis predicted a difference in selection pressure between cases and controls. In viral populations, selection pressure acts on the proteins and the genetic elements possessed by the virus. Owing to viral replication speeds, genetic mutation is likely to be the viruses' main mechanism for immune escape. Therefore, any differences between active and inactive disease are likely to be evidenced at the genetic level.

This chapter is structured into three sub-sections: alignments, phylograms, and diversity.
The initial section on alignments details the use of the alignments to remove any nonsense mutations. The following section on phylograms displays the phylogenetic trees generated from our alignments, and includes the confirmation of maximum likelihood. Finally the section on diversity data covers the pairwise diversity, the genetic composition, and the location and number of changes seen.

During these analyses, we specifically focused on any emerging patterns that differed between cases and controls. Given our hypothesis, we expected to find less genetic diversity within 'case patients' (active disease), when compared to 'control patients' (inactive disease). This would infer a reduced level of immune activity within active disease.

It is important to note that in the following sections, "active disease" refers to viruses extracted from patients with active disease (e-CHB, 'cases'); and "inactive disease" refers to viruses extracted from patients with inactive disease (e-InD, 'controls'). The terms are used interchangeably.

3.2 Materials and Methods

3.2.1 Materials

69 HBV (genotype C & D) eAg+ clone sequences, length 1,125bp (a generous gift from Dr. Bill Abbott) and 127 HBV (genotype C & D) eAg- clone sequences obtained by the methods described above. *Total number of sequences: 196*

3.2.2 Sequence Analysis

Sequence chromatograms were obtained using the ABIPOP-3000 courtesy of Kristine Boxen, Centre for Genomics and Proteomics, Auckland University. These were then imported into the bioinformatic analysis application Geneious©, Courtesy of Biomatters, Ltd[38]. All sequences were trimmed to begin at the primer-binding site, motif TGGTATTGCCCAAG. Sequences were edited manually, in accordance with the chromatogram to preserve all ORFs without premature stop codons, and correct any errors made by the sequencing instrument. In cases where sequence quality was poor, the sample was re-sequenced using a reverse primer, and a consensus sequence was generated from the contig of the two. The minimum overlap was set to 25, and the Overlap Identity was set to 80%. Gap Open / Extend Penalty was set to 18, Mismatch Score was set to -9, and Match Score was set to 5.

3.2.3 Trimming

The 1,125 amplimers generated from the sequencing assay were of variable quality, and thus required trimming to a consistent length. It was decided that 900bp retained enough genetic information, whilst removing the greatest amount of low quality sequence, which could distort any analyses.

3.2.4 Patient Specific Alignment and Tree Building

All clones from each patient were aligned (nucleotide and amino-acid) and checked for insertions and deletions leading to misalignment, in order to ascertain sequence integrity. These alignments were then used to generate an unrooted Neighbour-Joining (NJ) tree to elucidate outliers and anomalies, which were then excluded. These trees were created using the following parameters: Cost Matrix 65% Similarity; Gap open penalty 12; Gap extension penalty 3; Global Alignment; Genetic Distance Model HKY; No Out-group. Average pairwise diversity was also calculated using Geneious[38].

3.2.5 Extractions

Once sequence quality and integrity was confirmed, and anomalies removed, four different replicates of the dataset were created. 1) A “whole amplimer” dataset [900bp]; 2) A “Core Gene only” dataset [552bp - restricted to the core ORF]; 3) An “eAg and Core” dataset [638bp - restricted to the core ORF and the e ORF]; 4) An “eAg only” dataset [95bp - restricted to the e ORF].

The alignment from Geneious[38] was then imported into PAUP*[39], which computes a maximum-likelihood phylogenetic tree based on amino acid differences. The results of both methods were then compared to check for convergence. (i.e. - similar topologies)

3.2.6 Data Set alignments and Tree construction

Each data set was aligned by Clustal-W[40], using the cost matrix “IUB”; gap open penalty “15”; gap extension penalty “6.66”.
PhyML[41] was used to generate Maximum Likelihood trees from these alignments.

3.2.7 Shimodaira and Hasegawa Test

The SH test is used to compare the likelihood of the given ML tree with a predefined ‘constrained tree’. Using PAUP*, we generated a tree which constrained the topology for a single patient to be monophyletic, and compared this to the unconstrained ML tree generated using PhyML. This was repeated for each patient. This test yields a p-value that indicates whether the artificially constrained tree is of statistical significance, and thus whether the ML tree is representative of the ‘real tree’.

3.2.8 Pairwise comparison

Four E antigen positive clones were selected at random from the data set, and a consensus sequence was generated to establish a baseline. All clones from each patient were aligned and then compared to this baseline for deviations (i.e. - mutations) within the core gene. If a patient contained ≥ 1 deviations from the baseline, this was recorded as “true”; no deviations were recorded as “false”. This was repeated for each patient individually, and for the translated data set.

3.2.9 MacClade

MacClade is phylogenetic software that provides an interactive environment for exploring phylogeny, whereby phylogenetic trees can be manipulated and character evolution visualized upon them. MacClade contains a useful ‘chart’ function, which allows the mapping of an alignment to the corresponding phylogenetic tree, and calculates the region on the tree where changes occurred (Character evolution). This can be analysed in a variety of fashions, including the number of changes per site, the number of sites that contain a certain number of changes, or the frequency of changes at codon position 1, 2 or 3.

To visualise the character evolution occurring within each patient, we exported each alignment and corresponding ML tree from Geneious as a nexus file, which was then imported into MacClade for analysis. The above ‘chart’ function was then employed to analyse the frequency, location, and number of changes occurring.

3.3 Results

3.3.1 Alignments

Frequency of nonsense mutations

The insertion or deletion of a nucleotide character(s) can have deleterious effects on the virus, as this will alter the reading frame of the translation machinery, often producing a truncated protein. The alignment produced by Clustal-W was used to elucidate any clones containing nonsense mutations. These clones are unlikely to be derived from viable, infectious virus. Furthermore, the phylogenetic software used for these analyses cannot process alignments that contain gaps. Therefore, these clones were excluded from all further analyses.

To compare the frequency of nonsense mutations in cases and controls, the number of non-viable clones was divided by the total number of clones extracted. A signed rank test calculated a p-value of 0.19. The data are shown in Table 3:1, below.

This indicates that there is no significant difference in the frequency of non-sense mutations between cases and controls.

Table 3:1 - Signed rank test of nonsense mutations (p= 0.19)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.0833	0.5	0.125	0	0.2083	0.125
2	0.0769	0.7692			0.7692	0.6923
3	0.1428	0.1428	0.6		0.3714	0.2285
4	0	0.0833			0.0833	0.0833
5	0	0			0	0
6	0.1	0			0	-0.1

3.3.2 Phylograms of complete core gene clones:

Phylograms display the evolutionary history of an individual. Using Geneious and PhyML, we constructed a maximum likelihood tree of the core ORF for 196 clones (Figure 3:1).

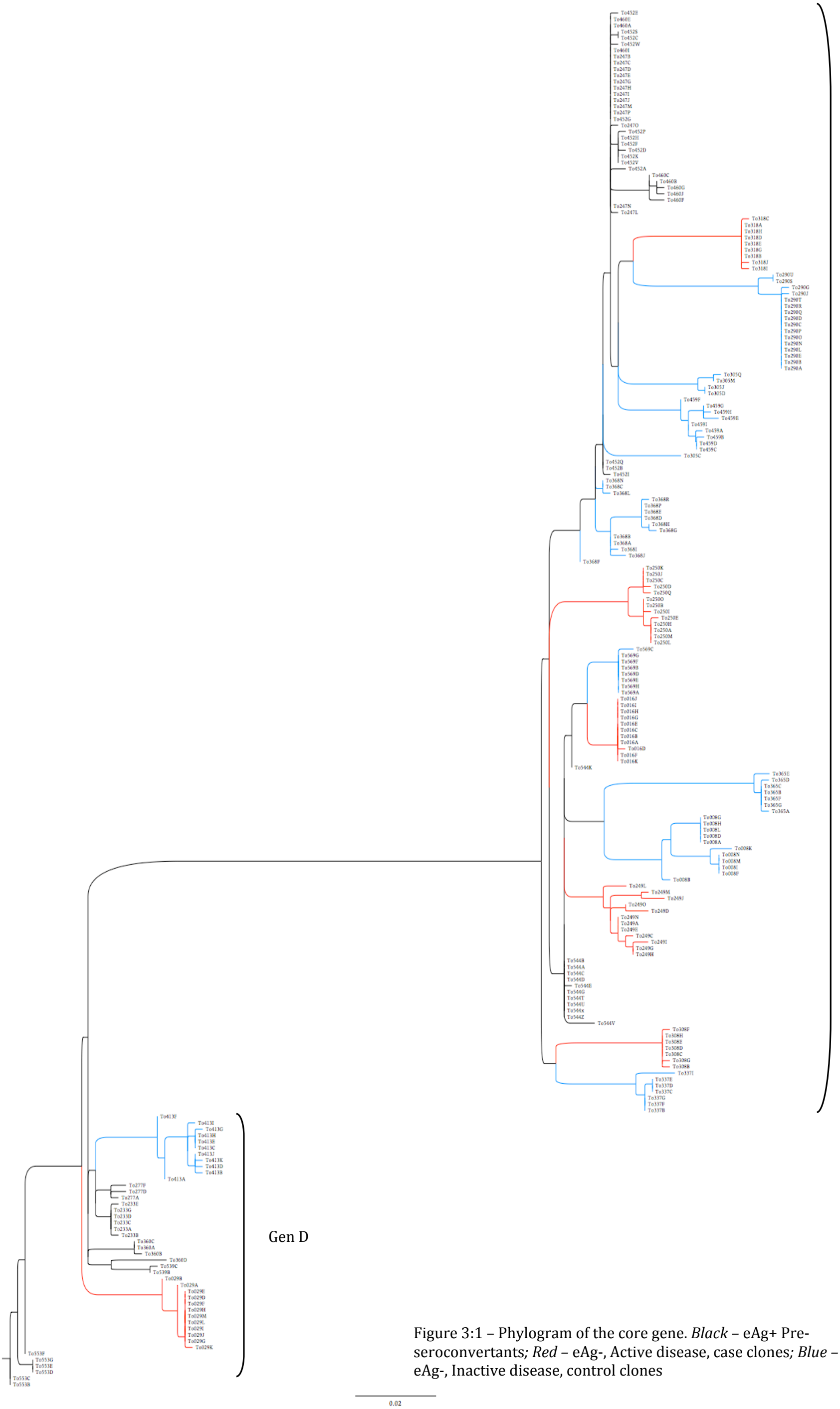
As expected, and consistent with previously described results [42], all HBeAg+ subjects had short external branches, whilst all HBeAg- subjects had long external branches. The long external branches are thought to be a result of HBeAg seroconversion. HBeAg+ subjects 460 and 360 are exceptions. It is speculated that these patients may be undergoing seroconversion.

Figure 3:1 also demonstrates two major arms, representing the HBV sub-genotypes C3 and D4, that are know to predominate in the Pacific Islands[33, 43]. The tree also demonstrates two further sub-sub-genotypes (C3a and C3b), first described by Lim et al in 2007[42]. These two sub-sub-genotypes are distinguished by four synonymous changes in the core gene, found at nucleotides 36, 87, 229, and 255.

The phylogram also demonstrates that all subjects were monophyletic, as previously reported[42], with the exception of subject 368. Subject 368 is a known recent HBeAg seroconvertant, and clustered near the top of the tree. Seroconversion is known to have occurred between 2000 and 2006. An examination of clinical data revealed that subject 368 was classified as HBeAg- in 2008, and again in 2009, therefore this observation is genuine. Furthermore, subject 368 demonstrates a high frequency of nonsense mutations (see appendix), which is consistent with a recent seroconversion.

Subject 553 is HBeAg -ve, and yet clustered at the top of the tree. Further examination revealed that no mutations were present in subject 553. Subject 553 however displayed a typical genotype D nucleotide sequence[44].

Subjects 290 (control) and 318 (case) demonstrated a recent divergence, indicating similarity. It is thought that this indicates that any immune deficit is host-specific, rather than virus-specific.



Gen C

Gen D

Figure 3:1 – Phylogram of the core gene. *Black* – eAg+ Pre-seroconvertants*; Red* – eAg-, Active disease, case clones*; Blue* – eAg-, Inactive disease, control clones

Verification of phylograms

The Shimodaira and Hasegawa test was used to ascertain whether the topology of the ML tree is representative of the 'real' tree. A p-value less than 0.05 indicates that there is a statistically significant difference between the ML tree and the constrained tree. The p-values obtained were adjusted using the Bonferroni correction. The result of this test (as shown in Table 3:2 below) is non-significant, indicating that the tree with the highest likelihood should be accepted. Therefore, the constrained tree was rejected and the ML tree (Figure 3:1) was accepted.

Table 3:2 – Shimodaira and Hasegawa test results, showing pre-corrected and post-corrected p-values.

Patient	Phenotype	P-Value	Adj. P-Value
008	Inactive disease	0.009	0.207
016	Active disease	0.009	0.207
029	Active disease	0.009	0.207
233	eAg+, Pre-seroconvertant	0.009	0.207
247	eAg+, Pre-seroconvertant	0.009	0.207
249	Active disease	0.009	0.207
250	Active disease	0.009	0.207
277	eAg+, Pre-seroconvertant	0.009	0.207
290	Inactive disease	0.009	0.207
305	Inactive disease	0.012	0.276
308	Active disease	0.009	0.207
318	Active disease	0.009	0.207
337	Inactive disease	0.009	0.207
360	eAg+, Pre-seroconvertant	0.008	0.184
368	Inactive disease	0.019	0.437
413	Inactive disease	0.009	0.207
452	eAg+, Pre-seroconvertant	0.008	0.184
459	Inactive disease	0.009	0.207
460	eAg+, Pre-seroconvertant	0.014	0.322
539	eAg+, Pre-seroconvertant	0.009	0.207
544	eAg+, Pre-seroconvertant	0.007	0.161
553	Inactive disease	0.009	0.207
569	Inactive disease	0.009	0.207

3.3.3 Analyses of Pairwise Diversity using Geneious

Pairwise diversity measures the level of diversity found within a monophyletic clade. Geneious[38] was used to calculate pairwise diversity within each subject. Overall diversity within subjects was low. The range for the Core gene was 0.00% - 1.60%. The range for the entire core ORF (including the eAg) was 0.10% - 1.40%. The range for the whole amplimer was 0.10% - 1.60%.

This data is shown in Table 3:3 Table 3:5, and a signed rank test was performed to determine if there is any difference between case and controls. The p-values of 0.22, 0.19, and 0.06 (respectively) were calculated.

It can be seen that a trend for increased diversity within controls was observed in all three data sets, and that a lack of significance was due to case subject 249 (group 3), who is an outlier in the case group.

Table 3:3 - Signed rank test of diversity within the core gene (p=0.22)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.00%	0.30%	0.40%	0.10%	0.0026	0.0026
2	0.10%	1.60%			0.016	0.015
3	1.00%	0.30%	0.80%		0.0055	-0.0045
4	0.60%	1.10%			0.011	0.005
5	0.20%	0.50%			0.005	0.003
6	0.10%	0.40%			0.004	0.003

Table 3:4 - Signed rank test of diversity within the core gene, including the eAg (p=0.19)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.10%	0.30%	0.30%	0.10%	0.0023	0.0013
2	0.20%	1.40%			0.014	0.012
3	1.00%	0.30%	0.90%		0.006	-0.004
4	0.60%	1.30%			0.013	0.007
5	0.10%	0.50%			0.005	0.004
6	0.10%	0.40%			0.004	0.003

Table 3:5 - Signed rank test of diversity, whole amplimer (p=0.06)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.10%	0.20%	0.40%	0.20%	0.0026	0.0016
2	0.30%	1.20%			0.012	0.009
3	0.80%	0.20%	1.20%		0.007	-0.001
4	0.80%	1.60%			0.016	0.008
5	0.20%	0.40%			0.004	0.002
6	0.10%	0.30%			0.003	0.002

Genetic Composition

Using Geneious, average nucleotide and amino acid composition was calculated for cases and controls separately. The data are shown in Figure 3:2 Figure 3:3. There is no significant difference between cases and control genetic composition.

Nucleotide Composition

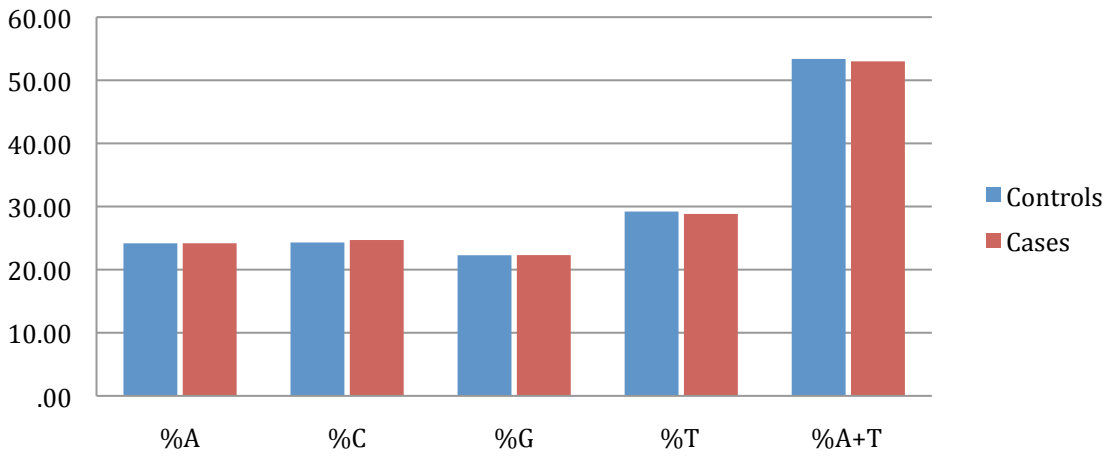


Figure 3:2 - Average Nucleotide composition (%) for cases and controls.

Amino Acid Composition

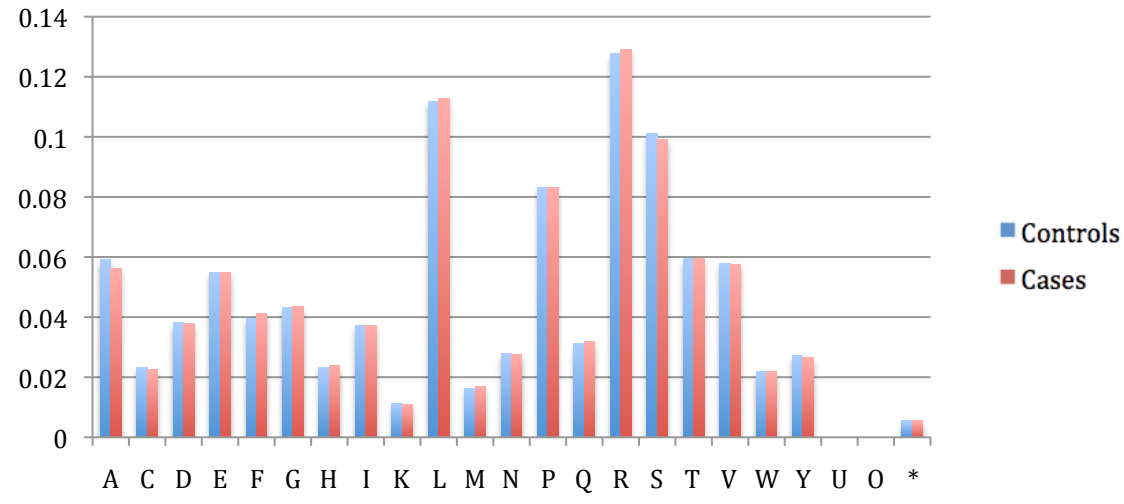


Figure 3:3 - Amino acid composition of cases (red) and controls (blue)

3.3.4 Analyses of Diversity using MacClade

Codon position analysis:

MacClade was used to tabulate the average location of nucleotide changes within subjects (position 1, 2, or 3). Changes at position 1 result in a residue change in most instances (i.e.- a non-synonymous change). Changes at position 2 may or may not result in a residue change. Due to the redundancy of the triplet nucleotide code, changes at position 3 in most instances do not result in a residue change (i.e. – a synonymous change).

A signed rank test was conducted on the number of first and second position changes (Table 3:6) and the number of third position changes (Table 3:7). Both results were non-significant (p=0.34, p=0.88; respectively).
This indicates that there are no significant differences in codon position changes between cases and controls.

Table 3:6 - Signed rank test of the number of first and second position changes (p=0.34)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	0	2	1	1	1
2	3	16			16	13
3	13	1	11		6	-7
4	7	6			6	-1
5	0	8			8	8
6	2	4			4	2

Table 3:7 - Signed rank test of third position changes, (p=0.88)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	1	4	2	1	2.3	1.3
2	1	4			4	3
3	11	6	4		5	-6
4	2	2			2	0
5	3	3			3	0
6	1	3			3	2

Character changes:

MacClade allows the tracking of user-defined 'characters' across the phylogram. In our analysis, we defined the characters as amino acids. We used this function to calculate the number of times each amino acid within the core gene changed. This allowed the highlighting of any amino acids that changed once or more.

A signed rank test was performed on the number of sites that demonstrated a single change (Table 3:8) and the number of sites that demonstrated a double change (Table 3:9).
A p-value of 0.16 was calculated for the number of single changes. A p-value of 1.0 was calculated for the number of double changes. This indicates that there is no significant difference in number of character changes between cases and controls.

Table 3:8 - Signed rank test of the number of individual sites displaying a single change (p=0.16)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	0	2	1	1	1
2	1	15			15	14
3	6	1	8		4.5	-1.5
4	4	4			4	0
5	0	6			6	6
6	2	4			4	2

Table 3:9 - Signed rank test of the number of individual sites displaying a double change (p=1.00)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	0	0	0	0	0
2	0	0			0	0
3	4	0	1		0.5	-3.5
4	0	2			2	2
5	0	1			1	1
6	0	0			0	0

Nucleotide diversity within clades

MacClade was used to elucidate the number of substitutions that were occurring within subjects. A signed rank test was performed (Table 3:10), which revealed a non-significant p-value of 0.63.

Therefore, there is no statistically significant difference in nucleotide changes between cases and controls.

Table 3:10 - Signed rank test of number of nucleotide changes within clades (p=0.63)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	1	4	2	1	2.3	1.3
2	2	4	7		5.5	3.5
3	23	7	13		10	-13
4	8	8			8	0
5	3	9			9	6
6	3	6			6	3

Number of non-synonymous changes

MacClade was used to elucidate and calculate the location of all non-synonymous (NS) amino acid changes. These are shown in the tables below

Group 1

Patient	Residues
16	
365	72
553	146
569	155

Group 2

Patient	Residues
29	21
305	24, 94
459	21, 27, 48, 50

Group 3

Patient	Residues
249	13, 74, 77, 87, 116, 154, 167
290	77
368	29, 31, 50, 103, 107, 130, 142

Group 4

Patient	Residues
250	43, 91, 179
8	83, 105, 181, 182

Group 5

Patient	Residues
308	
413	12, 21, 22, 36, 80, 113, 156

Group 6

Patient	Residues
318	48, 168
337	116, 165, 182

A signed rank test was performed to compare the number of non-synonymous changes between cases and controls (Table 3:11, below). This revealed a non-significant p-value of 0.34

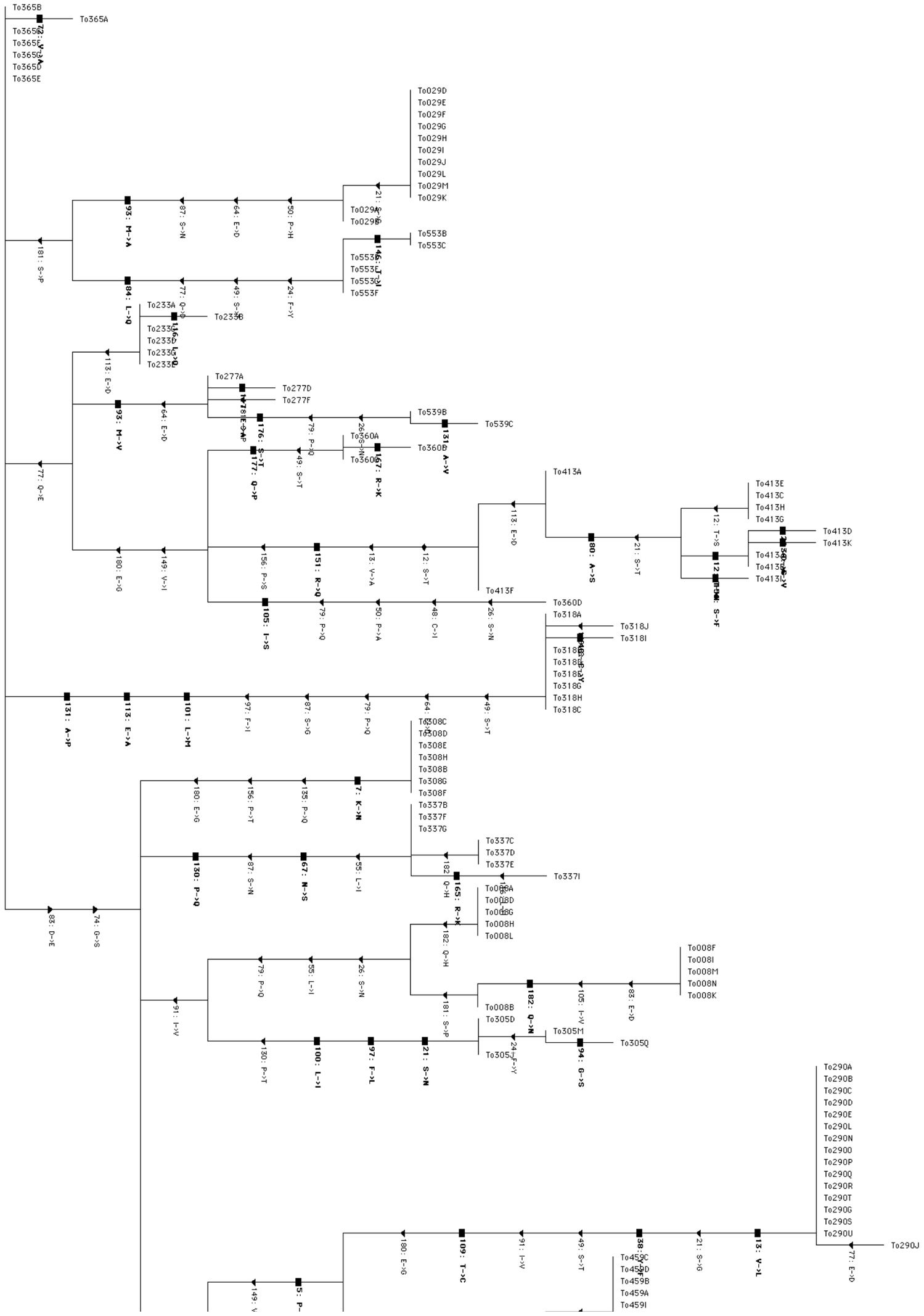
Table 3:11 - Signed rank test of number of non-synonymous changes (p=0.34)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	1	1	1	1	1
2	1	2	4		3	2
3	8	1	6		3.5	-4.5
4	3	5			5	2
5	0	8			8	8
6	2	3			3	1

MacClade Amino acid changes

MacClade was used to generate a phylogram showing the location (within the phylogram) of all amino acid changes in the core gene.

This is shown in Figure 3:4, p53.


To365B
To365A
To365C
To365D
To365E
To029D
To029E
To029F
To029G
To029H
To029I
To029J
To029L
To029M
To029K
To029A
To029B
To553B
To553C
To553D
To553E
To553F
To233A
To233B
To233C
To233D
To233G
To233E
To277A
To277D
To277F
To539B
To539C
To360A
To360B
To413A
To413E
To413C
To413H
To413G
To413D
To413K
To413F
To360D
To318A
To318J
To318I
To318C
To308C
To308D
To308E
To308H
To308B
To308G
To308F
To337B
To337F
To337G
To337C
To337D
To337E
To337I
To008A
To008D
To008G
To008H
To008L
To008F
To008I
To008M
To008N
To008K
To008B
To305D
To305M
To305Q
To305
To290A
To290B
To290C
To290D
To290E
To290L
To290N
To290O
To290P
To290Q
To290R
To290T
To290G
To290S
To290U
To290J
To459C
To459D
To459B
To459A
To459I

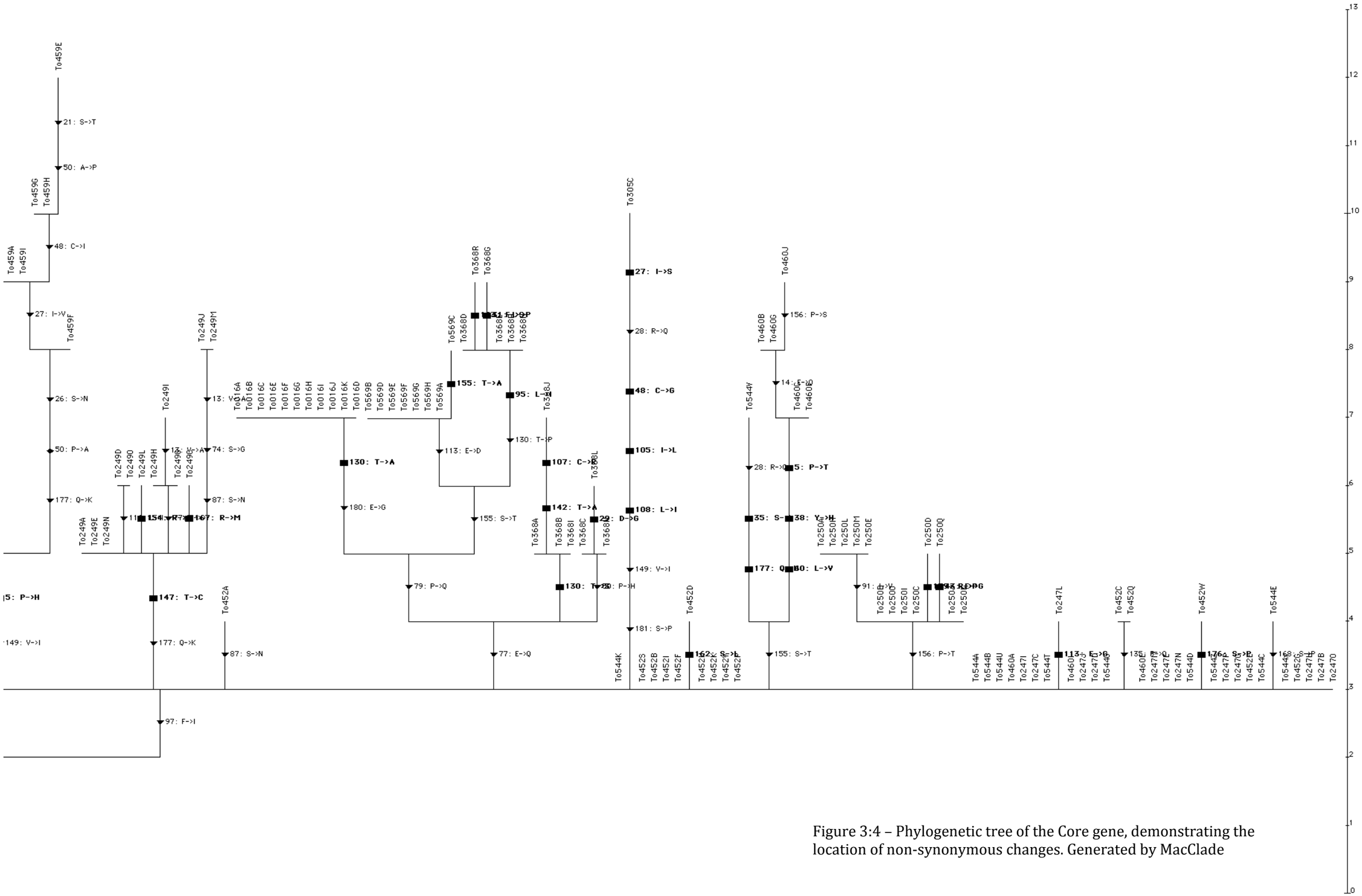


Figure 3:4 – Phylogenetic tree of the Core gene, demonstrating the location of non-synonymous changes. Generated by MacClade

Summary

An analysis of the alignments, phylograms, and diversity within the dataset did not demonstrate a statistically significant difference between active and inactive disease. However, as Table 3:12 shows, a strong trend towards a subtle deficit in active disease (case) subjects is observed. Furthermore, case patient 249 (group 3) is consistently an outlier in all analyses performed.

Table 3:12 - Summary of parameters used to compare diversity in cases and controls, showing the associated p-values

Analysis	Case	Control	p-value
Frequency of nonsense mutations	0.40	1.43	0.19
Core gene diversity	0.020	0.0441	0.22
Core ORF including eAg diversity	0.021	0.0443	0.19
Whole amplimer diversity	0.023	0.0446	0.06
Number of non-synonymous changes	14	23.5	0.34
Number of first and second position changes	25	41	0.34
Number of third position changes	19	19.3	0.88
Number of single amino acid changes	13	34.5	0.19
Number of double amino acid changes	4	3.5	1.00
Number of changes within clades	40	40.83	0.63

3.4 Discussion

Chapter 3 focussed on the alignments, the phylograms, and the diversity. This will be discussed here.

3.4.1 Alignments

We used Clustal-W to generate alignments of the cloned DNA sequences, for the purpose of removing nonsense mutations. Nonsense mutations are unlikely to be derived from a viable, infectious virus, and will cause disruptions to software packages such as PAML. We found 26 clones that demonstrated a nonsense mutation. A signed rank test showed no significant difference between the frequency of nonsense mutations in active and active disease.

Examination of the number of *patients* demonstrating deleterious clones (as opposed to the number of *clones*) shows an equal occurrence, indicating that this effect is most likely due to sample size. The similar rates found in pre-seroconvertant patients (data not shown) indicate that these observations are likely to be stochastic.

In addition, subject 305 consistently demonstrated deleterious mutations, most often presenting as a premature stop codon. This presented as a constant difficulty in this study, with only four of 18 clones not containing the premature stop codon.

3.4.2 Phylograms

We used Geneious and PhyML to construct a maximum likelihood tree of the data. The resultant phylogram was consistent with previously described characteristics of HBV evolution. Firstly, HBeAg +ve clones generally demonstrated short external branches, indicating no mutations had taken place; while HBeAg –ve clones demonstrated long external branches, which is thought to be a result of HBeAg seroconversion. This is consistent with results reported by Akarca et al in 1995[8]. Secondly, all subjects were monophyletic, as previously described by Rodrigo et al[45], with the exception of subject 368, who is a recent seroconvertant. Thirdly, previously identified HBV genotypic and sub-genotypic differences are corroborated by this phylogram[43]. Fourthly, the tree is consistent with a host-specific deficit.

The Shimodaira and Hasegawa test also confirmed the validity of this tree.

3.4.3 Analyses of Pairwise Diversity using Geneious

Using Geneious, an examination of the diversity found within the data demonstrated a strong trend towards a significant difference, with the whole amplimer having a p-value of 0.06. As the region under examination increased from 552bp to 900bp, the diversity also increased. No significant difference was noted in genetic composition.

In the absence of other factors, such as pharmacological treatments, a high degree of diversity within a clade is taken to infer one of two things: Either, the patient is demonstrating multiple infections; or alternatively the virus is being forced to mutate to survive. More accurately, only those viral clones that contain certain mutations survive (purifying selection). If the latter is true, this is seen as evidence of immune activity within the patient. Conversely however, the lack of diversity cannot be taken to infer that the immune system is inactive.

This is due to several factors. The timing of the sampling will have an effect on the diversity observed. Consider an infection that has persisted in the host for 24 months compared to one which has persisted for 96. Logically, the longer the persistence (length of infection) the greater the chance of selection creating a high

number of quasi-species, and thus a higher diversity. Thus a sample taken early in the infection will exhibit less diversity than a sample taken at a later time point. The length of infection is difficult to estimate because it is impossible to ascertain the time point of infection, especially in endemic areas and populations, such as Tonga.

Given the possible effects of the time of sampling, clonal expansion may also play a part in altering the observed diversity. During the course of selection within the host, it is conceivable that a single viral clone may find (via mutation) an epitope 'niche' that is not recognised by the immune system. Due to viral replication occurring asexually, all viral progeny are clones of the parent, and they will continue to fill this niche, unabated. Thus the only diversity that occurs is through random mutation. This has the net effect of creating a viral population with low levels of diversity, due to high fixation. At a phylogenetic level, this will exhibit as a long external branch (representing the initial mutation(s)), and then a clade of short internal branches (representing random mutation). Therefore, our samples may contain *historic* changes, (representative of immune escape), but very few *current* changes. Thus if our time of sampling is post-immune escape, this is likely to exhibit as a low diversity *within* patients, as they are no longer experiencing the high levels of immune selection pressure.

An interesting further study would be to follow the diversity of clones within a patient longitudinally, as they progress from acute hepatitis through to active / inactive disease. The authors would expect that a graph of the inactive disease patients' diversity versus time would show a bell curve distribution.
Also of note is the observed phenomenon of compartmentalisation seen to occur in HIV. If such a phenomenon exists in HBV, this will have effects on the diversity seen.

Additionally, we found that doubling the number of clones sequenced from one subject did not influence the diversity measure (data not shown).

Furthermore, it can be seen that there are more patients with greater diversity in controls, than in cases (3 to 1, respectively). Although this is consistent with our hypothesis, it is a weak result, since it cannot be ruled out that this may be simply a product of sample size (controls, n=10; cases, n=6), and that the difference is

small. It is interesting to note however that the average number of clones sequenced for cases and controls was similar (10.33 and 9.1 respectively).

It is conceivable that a doubling of the sample size would increase the level of significance and further studies should consider using a greater number of patients.

This indicates that there is a subtle difference in genetic diversity between case and control patients, and that this is likely related to HLA haplotypes.

3.4.4 Analyses of diversity using MacClade

MacClade was used to analyse the location and number of changes found within the data.

Codon position analysis

Due to the redundancy of the genetic code, a nucleotide change will have different effects depending on its position in the codon. Most third position changes are benign, as they do not change the amino acid and therefore the tertiary and quaternary protein structure (synonymous change). Most first and second position changes however will alter the amino acid and therefore alter the protein (non-synonymous change). Non-synonymous changes indicate mutation, presumably to escape some form of selection. Within HBV this is assumed to be the immune system. Therefore, changes found at the first and second positions are indicative of selection pressure.

Examination of the location of changes within the codon shows that case clones demonstrate a lack of first and second position changes (Table 3:6, p47). However, this result was non-significant (p=0.34). A further examination of the number of non-synonymous changes within patients (Table 3:11, p50) also revealed a non-significant p-value (0.34).

A more detailed analysis of the location of changes within four subjects was conducted, and the results are included in the appendix (p150). Case patients demonstrate a low rate of first and second position changes (55 – 63% of all changes), whilst control patients demonstrate a higher rate (64 – 70% of all changes). Not all changes at first and second positions will result in a non-synonymous mutation, however these results are taken as indicative of subtly reduced immune activity in case patients.

Number of character changes:

MacClade allows the calculation of the number of times a single amino acid changes. Single changes are likely to be indicative of immune selection pressure. Double changes (whereby the same amino acid changes two times) are quite rare, but are shown to occur (Table 3:9). These are thought to be indicative of sites of importance. Since overall intra-patient diversity is low, most sites did not change (as expected).

Interestingly, cases demonstrate on average less single character changes than controls, with the exception of subject 249 (Table 3:8, p48). The number of double changes is equal between cases and controls.

3.4.5 Other considerations

Despite a trend towards a statistically significant difference in diversity between active and inactive disease found by Geneious, further detailed analyses into the location and nature of the changes seen using MacClade did not demonstrate a significant result.

An attractive caveat to this observation however exists. Examination of the data presented in chapter 3 reveals that case subject 249 (group 3) can consistently be classified as an outlier in all analyses performed. Therefore it stands to reason that if the effects of this outlier were mitigated by an increase in the number of matched pairs used, a significant result may be seen.

One attractive hypothesis is that subject 249 may be undergoing 'immune control' and thus is 'converting' from active to inactive disease. Conceivably, it is also possible that all inactive disease clones progressed through an active disease-like stage, albeit short-lived. This raises the possibility that perhaps the mechanism allowing rapid and efficient immune control of HBV is deficient in case subjects. If this hypothesis is correct, this then would require that at a certain point, case subjects would become classified as control subjects.

Examination of Table 2:3 (p27) reveals that subject 249 demonstrates the only occurrence of a liver disease classification of G4S4, indicating grade 4 (of 4)

lymphocyte infiltrate, and stage 4 (of 4) fibrotic scarring (cirrhosis). These observations are consistent with strong immune pressure on the hepatocytes, as a result of HBV infection.
Given that our study design was not longitudinal, these hypotheses cannot be tested, however our data does not exclude these possibilities.

The aim of this study was to create a therapeutic vaccine that would convert active disease to inactive disease. Further study is required to elucidate if patient 249 has undergone such a process, and then to examine the mechanisms behind this conversion.

3.4.6 Summary

In this chapter, we demonstrated that the sequence data obtained is consistent with previously described findings, most especially in relation to phylogram topology. In addition, we demonstrated a strong trend towards decreased diversity within active disease subjects, especially if subject 249 is considered to be an outlier. This trend was reinforced by further findings in the number and location of changes seen. The incidence of more mutations at positions 1 and 2 is consistent with increased selection pressure, although selection pressure studies need to be conducted to exclude the possibility of genetic drift. This is addressed in chapter 4..

In addition, further studies should examine the diversity in the other ORFs (Polymerase, X, and Surface).

Chapter 4:

Analysis of Selection Pressure

4 Analysis of Selection Pressure

4.1 Introduction

Our original hypothesis predicted an immune deficit within the active disease population, which allows the high levels of viral replication that are a feature of this disease. This may be due to reduced activity of either innate immunity or the CD8 T cell compartment of the acquired immune system. An investigation of selection pressure placed on the viral genome may give clues to the nature of the immune deficit that causes e-CHB (active disease).

In this chapter we start by conducting an analysis on the whole phylogenetic tree using PAML. First we looked for differences in selection pressure between active and inactive disease, and then we identified the specific amino acids in the core gene that were commonly under positive selection pressure. We then conducted analyses on individual patients using PAML. The purpose of these analyses was to confirm the results of the whole tree analyses. In addition this allowed the within-subject and between-subject components of the tree to be studied independently, which might give information about the different mechanisms that influence these components.

4.2 Materials and Methods

4.2.1 Materials

As in chapter 3, 69 HBV (genotype C3 & D4) HBeAg+ clone sequences, length 1,125bp (a generous gift from Dr. Bill Abbott) and 127 HBV (genotype C3 & D4) HBeAg- clone sequences obtained by the methods described in chapter 2. *Total number of sequences: 196*

NB: All analyses were performed on the core gene (552bp).

4.2.2 PAML (Phylogenetic Analysis by Maximum Likelihood)

PAML is a command line software package of programs used for phylogenetic analyses using maximum likelihood, and includes BASEML, CODEML, and Evolver.

CODEML can be used to calculate, among other things, the proportion of sites under selection (negative, neutral, positive), as well identifying the sites under selection and calculating the dN/dS ratio (Ratio of non-synonymous changes to synonymous changes, otherwise referred to as omega [ω]).

Inputs:

The Control File:

```
      seqfile = alignment.phy    * sequence data filename
    treefile = ML.tree           * tree structure file name
     outfile = Model_2a.out      * main result file name
```

As demonstrated above, PAML requires an alignment and a tree file. For our analyses the alignments were processed in PHYLIP (.phy) format, and the trees in newick format.

Models:
In our analyses we used the site models Nearly-Neutral ('M1a'), Positive-Selection ('M2a'), and the branch-sites model ('model A').

The Control File:

```
        model = 0,1,2

                * models for codons:

                      * 0:one, 1:b, 2: 2 or more dN/dS ratios
                      for branches

      NSsites = 0,1,2

                *0:one w; 1:neutral; 2:selection;
```

Model A: `[model=2, nsites=2]`

Model A is referred to as a 'branch-site' model and aims to detect positive selection on a designated branch. `model=2` allows omega to vary between branches, and `nsites=2` allows omega to vary between sites in the gene. The utilisation of this model requires the branch(es) of interest to be designated as 'foreground' by adding the 'tag' "`#1`" to the branch.
Model A allows for four classes of selection, in both the foreground and the background: Class 0, Negative ($\omega < 1$); Class 1, Neutral ($\omega = 1$); Class 2a, Positive ($0 < \omega_0 < 1$, $\omega_2 \geq 1$); Class 2b, Positive ($\omega_1 = 1$, $\omega_2 \geq 1$).

Model M1a: `[model=0, nsites=1]`

Model M1a is also referred to as the 'nearly neutral' model, and aims to elucidate the proportion of sites under negative or neutral selection, and assign an omega value to both groups. `model=0` allows for the same model of evolution for all codons, and `nsites=1` allows for sites to be classified as negative ($\omega < 1$) or neutral ($\omega = 1$).

Model M2a: `[model=0, nsites=2]`

Model M2a aims to elucidate the proportion of sites under negative, neutral and

positive selection, and assigns an omega value to each of these classes. Model M2a is also referred to as 'positive selection', as it lists which sites are under positive selection. `nsites=2` allows for sites to be classified as negative ($\omega < 1$), neutral ($\omega = 1$), or positive ($\omega > 1$).

4.2.3 Likelihood Ratio Test

It is necessary to ascertain if the positive selection found using model M2a is statistically significant compared to the results obtained from model M1a, which only tests for negative and neutral selection. This is achieved by the Likelihood ratio test demonstrated by:

```
|2*(LnL1-LnL2)|
```

Where: `LnL1` refers to the log likelihood of model M1a (generated by CODEML); and `LnL2` refers to the log likelihood of model M2a.

A Chi squared test is then performed on the result, using 2 degrees of freedom. This yields the p-value, which is then corrected using a Bonferroni correction.

4.2.4 Analysis of external selection

To ascertain whether the observed selection was occurring within patients or between patients, we constructed an ML tree of the *most ancestral clone* from each patient, using PhyML. [Substitution model: GTR; Transition/Transversion ratio; Estimated; Proportion of invariant sites: Estimated; Number of substitution categories: 4; Gamma distribution parameters: Estimated; Optimise topology: Yes; Optimise branch lengths and rate parameters: Yes;]
This tree was analysed for sites under positive selection using PAML [model M2a] and verified by a comparison to model M1a using a likelihood ratio test (LRT).

4.3 Results

4.3.1 Phylogenetic Analyses of the whole tree

PAML is a command line based program that can be used to compute the proportion of sites under positive, neutral and negative selection. In addition, PAML employs a Bayes Empirical Bayes (BEB) analysis to determine which sites are under positive selection.

Model A (Branch-Sites Model)

Model A tests for the level of selection occurring on a certain branch(es) (foreground) compared to the 'background' of the rest of the tree. Our first analysis designated the cases as the 'foreground' and the remainder of the tree as the 'background'. This was then repeated, designating the controls as the foreground.

The analysis of the case branches as the foreground showed minimal selection occurring. Figure 4:1 (p70) shows that 68% of sites were under negative selection, with the remaining (32%) being classed as neutral selection. Negative and neutral selection occurring on the background and foreground appears similar, given the identical omega values. No amino acids are found to demonstrate positive selection (2a and 2b).

The analysis of the control branches as the foreground showed some interesting results. When examining negative and neutral selection, both the foreground and the background show the same omega value, indicating the same dN / dS ratio. The proportion of sites under negative and neutral selection was 65% and 32% respectively (See Figure 4:2, p70)
Examining positive selection however shows a strong omega value (7.74) on the foreground branches and a low omega value (2a = 0.04; 2b = 1) on the background branches (See Table 4:1, p69). This indicates that on the foreground branches, ~3% of amino acids are experiencing strong positive selection, when compared to the background branches. This result therefore indicates that approximately six amino acids (3% of 184) are experiencing positive selection in controls, and not in cases.

Table 4:1 - Results from PAML branch-sites model (Model A), collective analysis. Omega is a measure of the ratio of non-synonymous to synonymous changes.

	Site Class	Negative Selection	Neutral Selection	Positive Selection (2a)	Positive Selection (2b)
Cases	*proportion*	0.68253	0.31747	0	0
	background omega	0.04161	1	0.04161	1
	foreground omega	0.04161	1	1	1

	Site Class	Negative Selection	Neutral Selection	Positive Selection (2a)	Positive Selection (2b)
Controls	*proportion*	0.64916	0.31782	0.02217	0.01085
	background omega	0.04047	1	0.04047	1
	foreground omega	0.04047	1	7.74491	7.74491

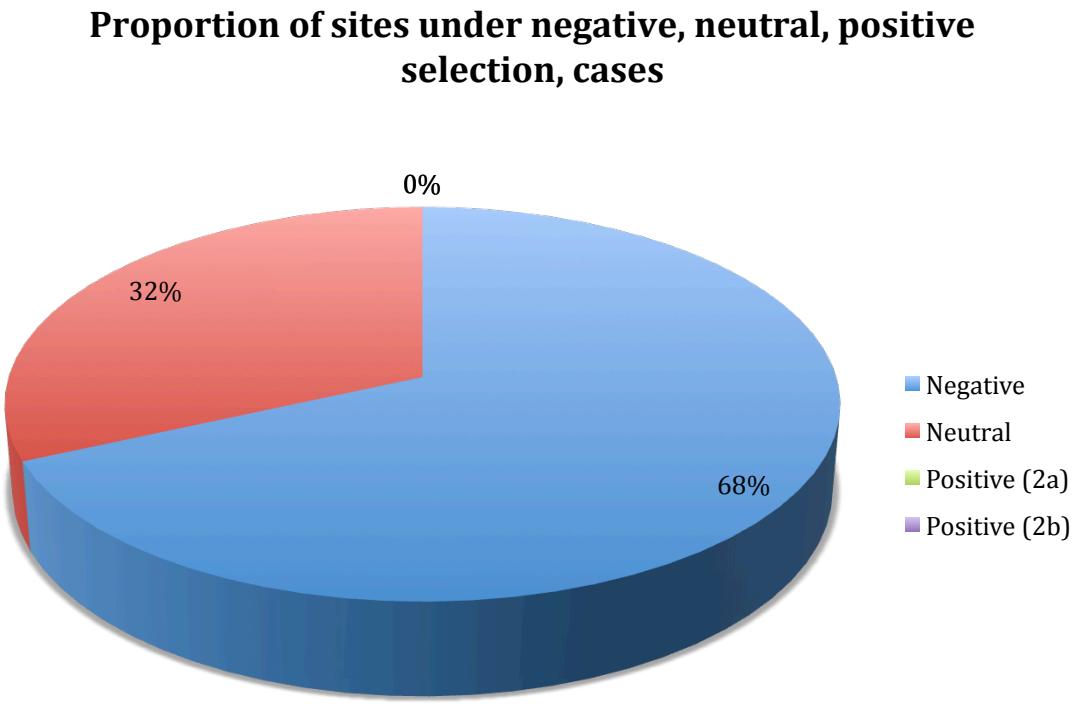


Figure 4:1 - Chart showing the proportion of sites under negative, neutral and positive selection when 'cases' are designated as a foreground. The chart demonstrates that under this analysis, 68% of sites are negatively selected, 32% are neutrally selected, and 0% are positively selected.

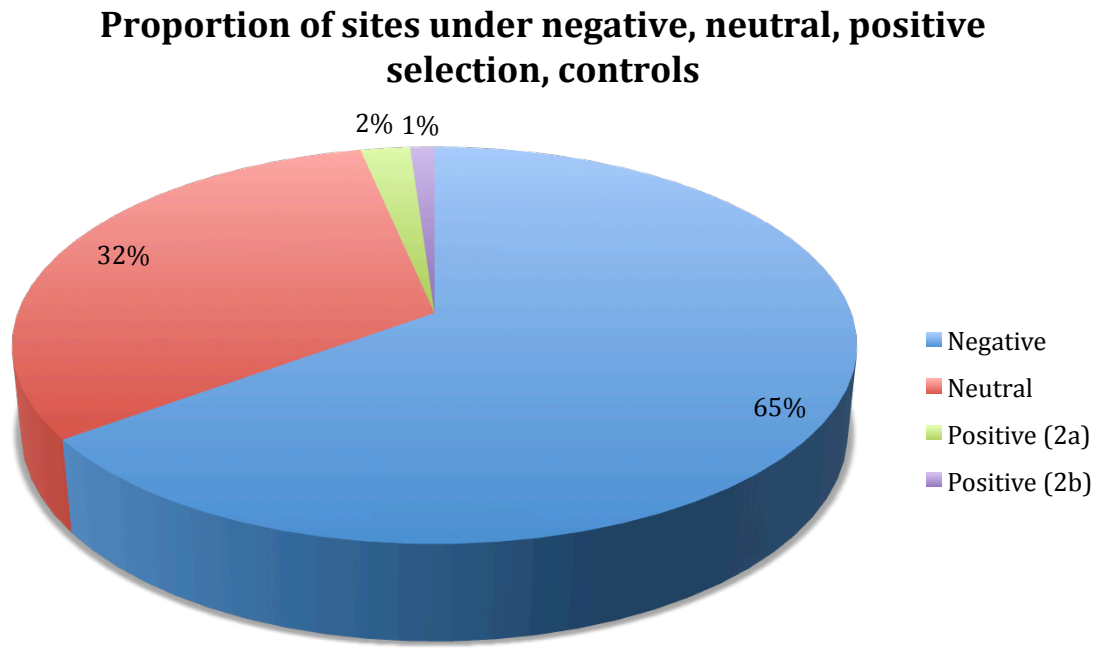


Figure 4:2 - Chart showing the proportion of sites under negative, neutral and positive selection when 'controls' are designated as a foreground. The chart demonstrates that under this analysis, 65% of sites are negatively selected, 32% are neutrally selected, and ~3% are positively selected.

Sites under positive selection in the whole tree

The Bayes Empirical Bayes analysis employed by PAML was used to elucidate sites under positive selection (Pr ≥ 0.50) within the whole tree. This was repeated with the Most Ancestral Clone, to check for consistency. The data are shown in Table 4:2 and Table 4:3, below. Six amino acids show strong positive selection (Pr ≥ 0.95).

Table 4:2 - Sites with a Pr ≥ 0.95 (p ≤ 0.05)

Site and Residue	Whole Tree		Most Ancestral Clone	
	Pr(w>1)	post mean +- SE for w	Pr(w>1)	post mean +- SE for w
21S	0.99**	3.795 +- 0.765		
26 S			0.98*	2.886 +- 0.832
77 E	0.99**	3.790+- 0.774	0.99*	2.910 +- 0.818
113 E	0.94	3.601 +- 0.989	0.95	2.813 +- 0.862
130 P	1.00**	3.796 +- 0.764	0.98*	2.904 +- 0.825
180 E	0.95	3.637 +- 0.957	0.95*	2.830 +- 0.859

Table 4:3 - Sites with Pr ≥ 0.50 (p ≤ 0.50)

Site and Residue	Whole Tree		Most Ancestral Clone	
	Pr(w>1)	post mean +- SE for w	Pr(w>1)	post mean +- SE for w
49 S			0.625	2.059 +- 0.930
64 E			0.609	2.021 +- 0.917
79 Q			0.72	2.250 +- 0.936
83 E			0.547	1.890 +- 0.884
87 S			0.73	2.257 +- 0.908
91 I			0.561	1.909 +- 0.872
109 T			0.74	2.292 +- 0.919
149 V			0.622	2.036 +- 0.901
151 R	0.595	2.592 +- 1.425	0.886	2.678 +- 0.932
177 Q			0.556	1.910 +- 0.891

Summary of analyses of full phylogenetic tree

In this section, we started by conducting an analysis on the whole phylogenetic tree, specifically aimed at elucidating differences in selection pressure between active and inactive disease subjects. The branch-sites PAML model was employed to compare the degree of selection occurring on the foreground (inactive disease) against the background of the remainder of the tree. This revealed that approximately six amino acids were under positive selection pressure ($\omega = 7.74$) in inactive disease subjects that were not found in active disease subjects.

PAML was then used to elucidate, via a Bayes Empirical Bayes analysis, which amino acids were commonly under positive selection. Six amino acids (21, 26, 77, 113, 130, 180) were highlighted, with $Pr \geq 0.95$. This was confirmed with a separate analysis of the Most Ancestral Clone. These sites are consistent with previously described sites under positive selection in Tongan subjects[45].

The findings from the branch-sites model are encouraging, as they are consistent with our hypothesis and demonstrate differential selection pressure between active and inactive disease. Both the proportion (2-3%) and the omega value (7.74) are within the expected range. Furthermore, the inverse analysis (active disease as the foreground) revealed no selection occurring, as expected.
It is however possible that the result from the branch-site model is a reflection of different sample sizes between active (n=6) and inactive (n=10) disease subjects. A positive value may be obtained simply due to more data being included in the foreground. Therefore, to obtain further evidence in support of this result, we conducted a paired study, using six HLA class-I matched groups.

4.3.2 Comparisons of selection pressure in HLA class I-matched groups

Our paired analysis of HLA class I-matched patients and controls was structured into two sections. First we analysed the frequency of non-synonymous mutations at the six positively-selected amino acid sites in individual patients, for the purpose of determining whether the controls had positive selection pressure at a greater number of these amino acid sites. We then conducted separate analyses of fixed (between-patient) and unfixed (within-patient) mutations. It is thought that fixed mutations are the result of strong selection pressure, whereas unfixed mutations may be the result of weak selection pressure. It is possible that this reflects two different mechanisms, thus requiring separate analyses.

4.3.2.1 - The frequency of non-synonymous mutations at positively selected amino acids.

To investigate if there were any differences in positive selection pressure between active and inactive disease subjects, we conducted an analysis of the number of patients who displayed a non-synonymous mutation at the six amino acids identified earlier in Table 4:2. A signed rank test revealed an almost significant p-value of 0.06. The data are shown in Table 4:4, below.

This indicates that inactive disease subjects demonstrate a strong trend towards an increased number of non-synonymous mutations at these six significant sites.

Table 4:4 - Signed rank test of the number of subjects containing a non-synonymous mutation at the six commonly positively selected amino acids (p=0.06)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	3.0	8.0	2.0	5.0	5.0	2.0
2	2.0	2.0	2.0		2.0	0.0
3	2.0	5.0	4.0		4.5	2.5
4	0.0	1.0			1.0	1.0
5	2.0	3.0			3.0	1.0
6	2.0	3.0			3.0	1.0

4.3.2.2 - Separate Analyses of Between-patient and Within-patient mutations.

A) Fixed Mutations (external)

1. Mutations occurring on external branches

MacClade was used to ascertain the number of changes occurring on external branches within the core gene.
Figure 4:3 (p75) shows the nine controls and the degree of changes occurring (indicated by shading). The high level of changes indicated by the black shading near the root of the tree is due to genotypic differences between the C3 and D4 clades. Subjects 290 and 365 demonstrate a high number of changes (12-15) from their most common ancestor. Figure 4:4 (p76) shows the same analysis repeated for the six cases. Subjects 250, 308, and 318 demonstrate a high level of changes (8-15) on external branches.

Control Patients

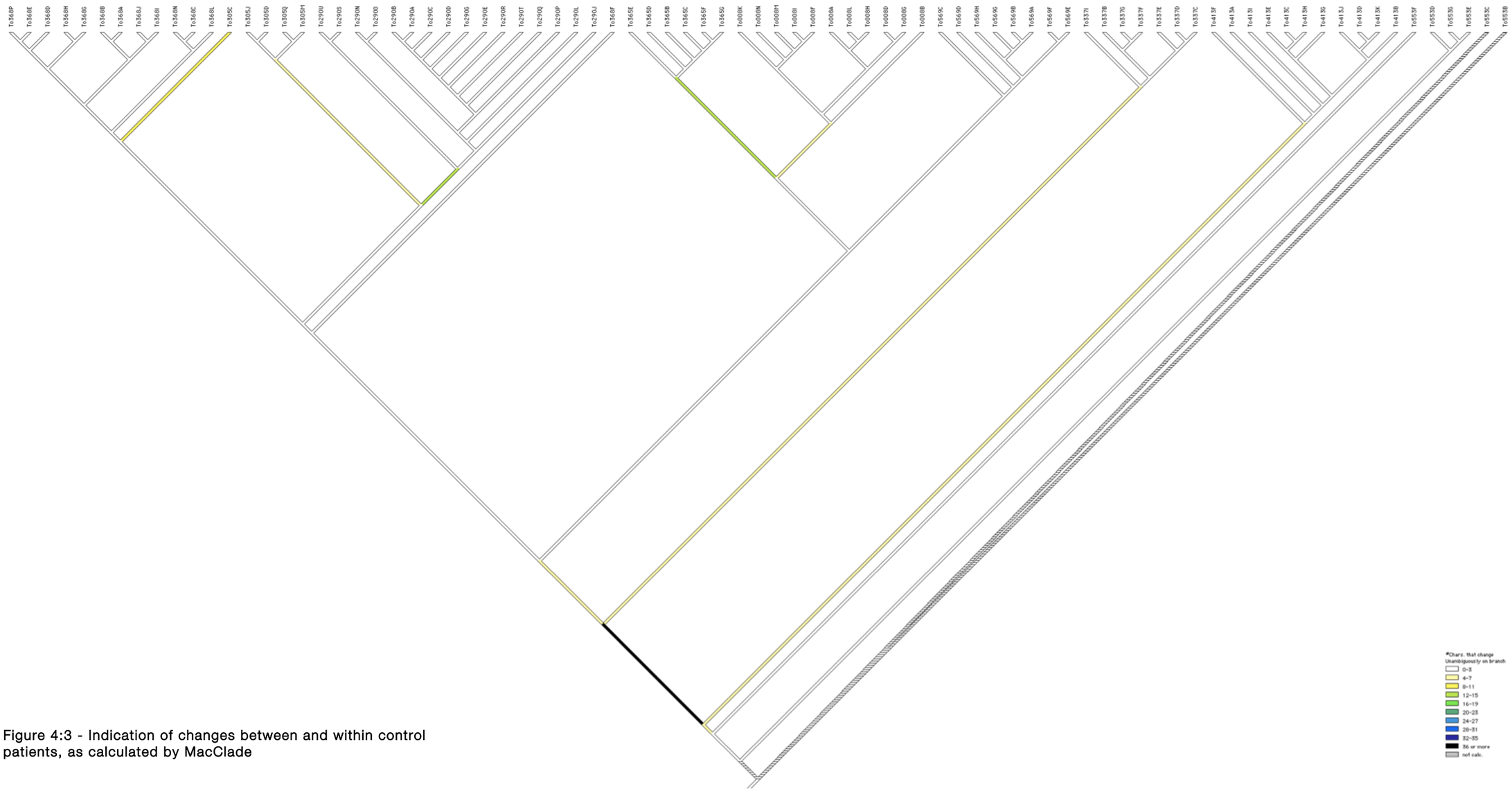


Figure 4:3 - Indication of changes between and within control patients, as calculated by MacClade

Case Patients

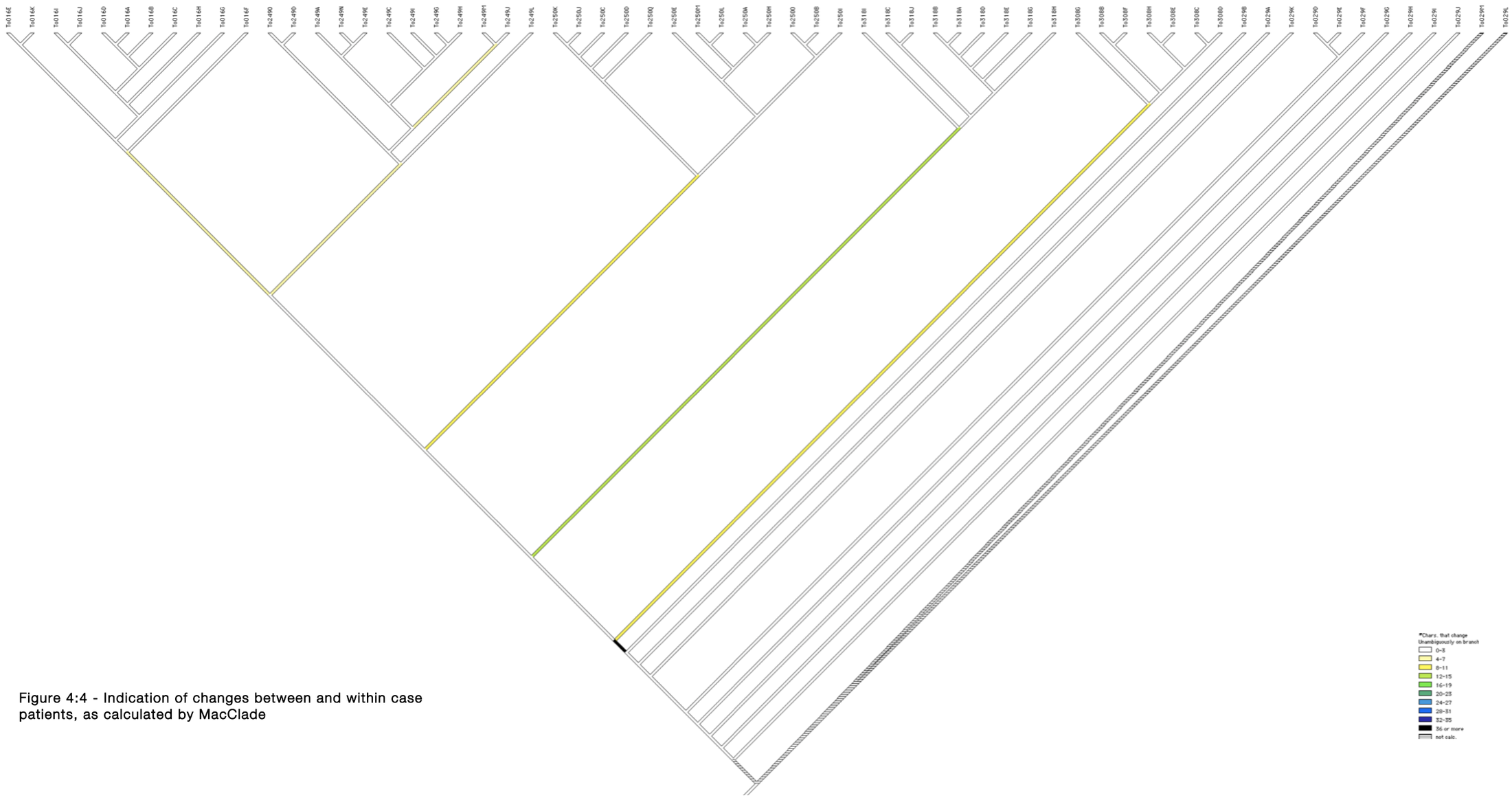


Figure 4:4 - Indication of changes between and within case patients, as calculated by MacClade

A signed rank test was performed on the number of changes found on external branches (Table 4:5, below). A non-significant result of p=1.00 was calculated.

Table 4:5 - Signed rank test of the number of non-synonymous changes occurring on external branches, (p=1.00)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	2	0	4	1	1.7	-0.3
2	4	4	3		3.5	-0.5
3	2	7	5		6	4
4	1	3			3	2
5	4	4			4	0
6	8	4			4	-4

2. Branch Length Analyses

Branch lengths reflect nucleotide changes. We compared the external branch lengths (as calculated by PAML) for active and inactive disease clades, by using the branch between the most recent divergence and the most ancestral clone. The external branch lengths were normalised by dividing their value by the total tree length (=0.82).

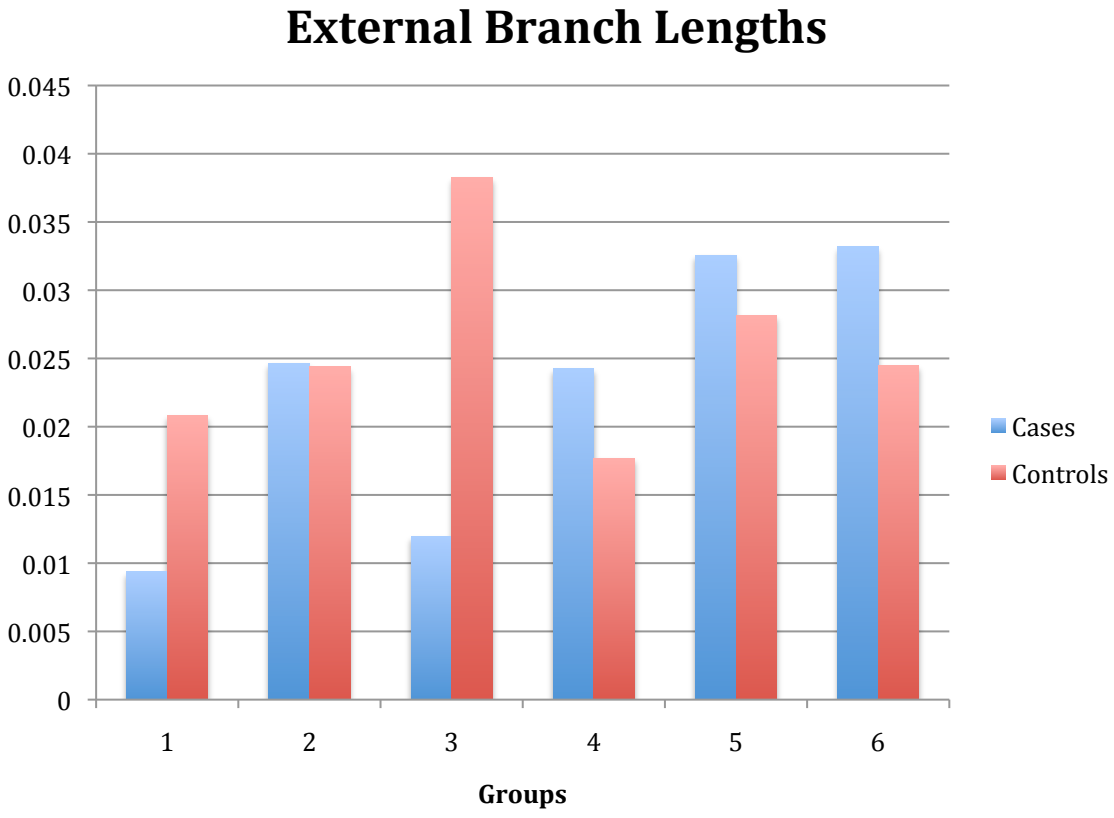


Figure 4:5 – External branch length analysis results, based on HLA class matching.

Branch lengths were surveyed for patterns as described in chapter 2 (Research rationale, p19). To refresh, if the active disease subjects demonstrated intermediate branch lengths (between HBeAg + baseline and inactive disease) this would indicate that there were fewer mutations in this cohort, which *could* infer that there was some immune deficit in the active disease cohort. This would be consistent with our hypothesis.

However, as can be seen in Figure 4:5 the active and inactive disease subjects do not exhibit an obvious difference in external branch lengths. For example it can be seen that there are both active and inactive disease patients who show long branch lengths, conversely there are also active and inactive disease patients who show short branch lengths.
In support of this, a signed rank test of external branch lengths was conducted, and found to be non-significant (p=1.0). The data are shown in Table 4:6, below.

Table 4:6 - Signed rank test analysis of external branches (p=1.0)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.0093	0.0459	0.0069	0.0095	0.0208	0.0114
2	0.0246	0.0243			0.0243	-0.0002
3	0.0119	0.0382			0.0382	0.0263
4	0.0242	0.0176			0.0176	-0.0065
5	0.0325	0.0281			0.0281	-0.0043
6	0.0331	0.0245			0.0245	-0.0086

3. PAML analysis of external selection

PAML model M2a was used to conduct an analysis on the levels of selection occurring on external branches, by creating an ML tree using only the most ancestral clone (MAC) from each clade.

The aim was to investigate selection occurring at sites *between* patients.

Figure 4:6 below shows the degree of negative, neutral and positive selection occurring on external branches.

External branches demonstrate a high proportion of sites under positive selection (8%). 22% of sites were classified as being under neutral selection, with the remaining 71% being classified as under negative selection.

The selection data were confirmed using the LRT. (See appendix)

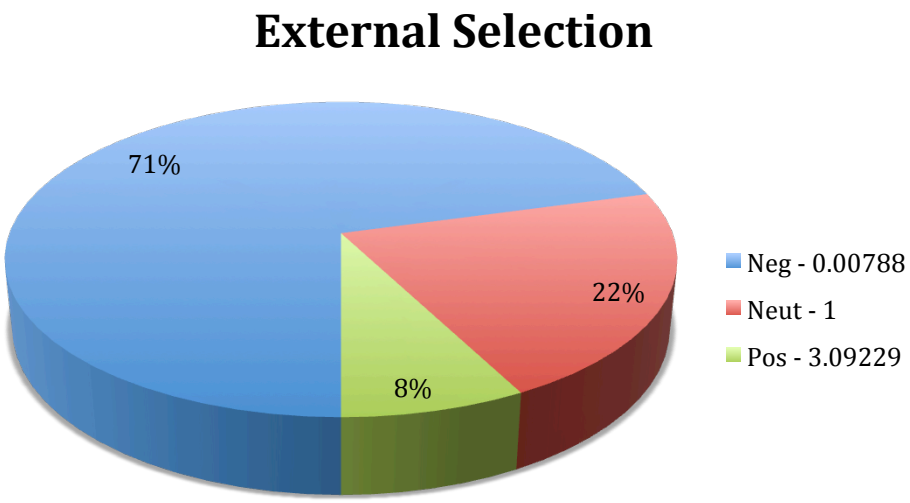


Figure 4:6 - Proportion of sites under selection on external branches (fixed mutations). *Blue* – Negative selection, 70.7%, ω = 0.0078; *Red* – Neutral selection, 21.7%, ω = 1; *Green* – Positive selection, 7.6%, ω = 3.092;

4. Sites under positive selection between patients

The PAML analysis of the most ancestral clone also highlighted, via a Bayes Empirical Bayes analysis, 14 sites that were under some form of positive selection between subjects. The data are shown in Table 4:7 and Table 4:8, below.

Table 4:7 - Sites with a Pr ≥ 0.95 (p≤0.05)

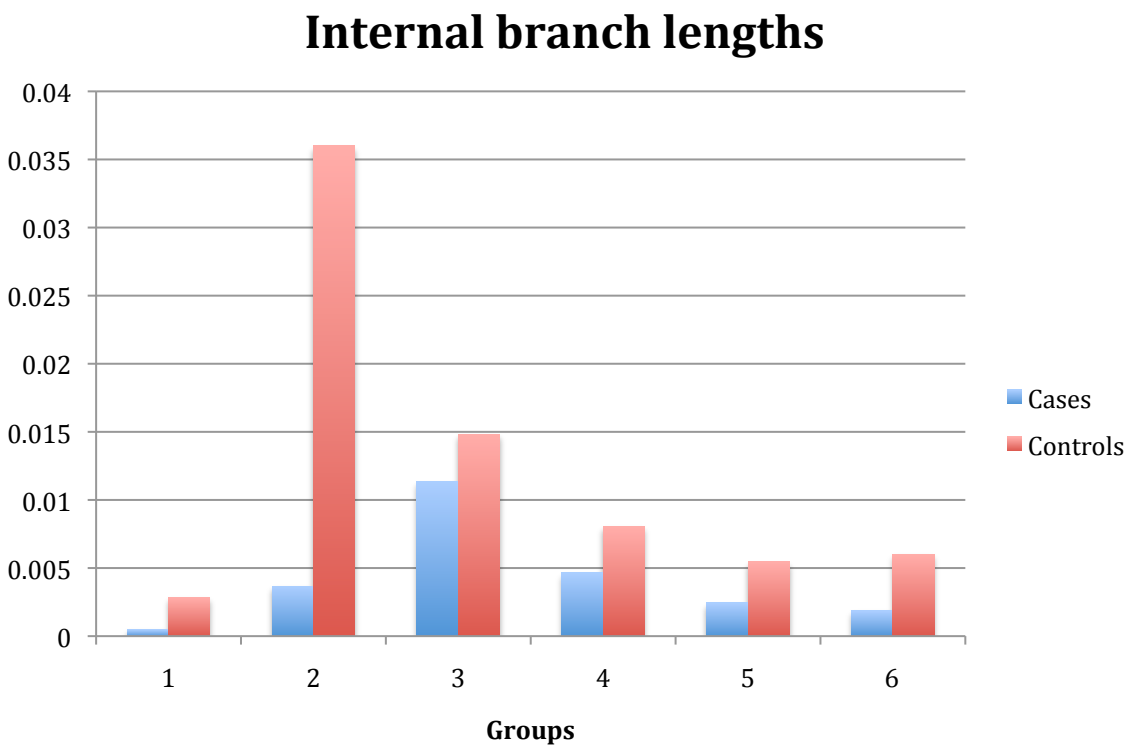
Site and Residue	Most Ancestral Clone	
	Pr(w>1)	post mean +- SE for w
26 S	0.98*	2.886 +- 0.832
77 Q	0.99*	2.910 +- 0.818
113 D	0.95	2.813 +- 0.862
130 T	0.98*	2.904 +- 0.825
180 E	0.95*	2.830 +- 0.859

Table 4:8 - Sites with Pr ≥ 0.50 (p ≤ 0.50)

Site and Residue	Most Ancestral Clone	
	Pr(w>1)	post mean +- SE for w
49 S	0.625	2.059 +- 0.930
64 E	0.609	2.021 +- 0.917
79 Q	0.72	2.250 +- 0.936
83 E	0.547	1.890 +- 0.884
87 S	0.73	2.257 +- 0.908
91 I	0.561	1.909 +- 0.872
109 T	0.74	2.292 +- 0.919
149 V	0.622	2.036 +- 0.901
177 Q	0.556	1.910 +- 0.891

B) Non-fixed Mutations (internal)

1. Branch Length Analyses



Branch lengths were retrieved from the PAML analyses, and normalised by the number of clones sequenced from that patient. A signed rank test (Table 4:9) calculated a statistically significant p-value of 0.03, indicating that branch lengths within case clones are significantly shorter than controls.

Table 4:9 - Signed rank test of internal branch lengths (p=0.03)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0.0005	0.0040	0.0032	0.0014	0.0028	0.0023
2	0.0036	0.0359	0.0359		0.0359	0.0323
3	0.0113	0.0066	0.0230		0.0148	0.0035
4	0.0046	0.0080			0.0080	0.0034
5	0.0024	0.0054			0.0054	0.003
6	0.0018	0.0060			0.0060	0.0042

2. Proportion of sites under positive selection within patients

We used PAML model 2a to elucidate the proportion of sites within each patient that were under negative, neutral, and positive selection.

Figure 4:7 - Figure 4:12 below display the proportion of sites under each class of selection. Table 4:10 (p85) displays the omega values calculated.

Group 1

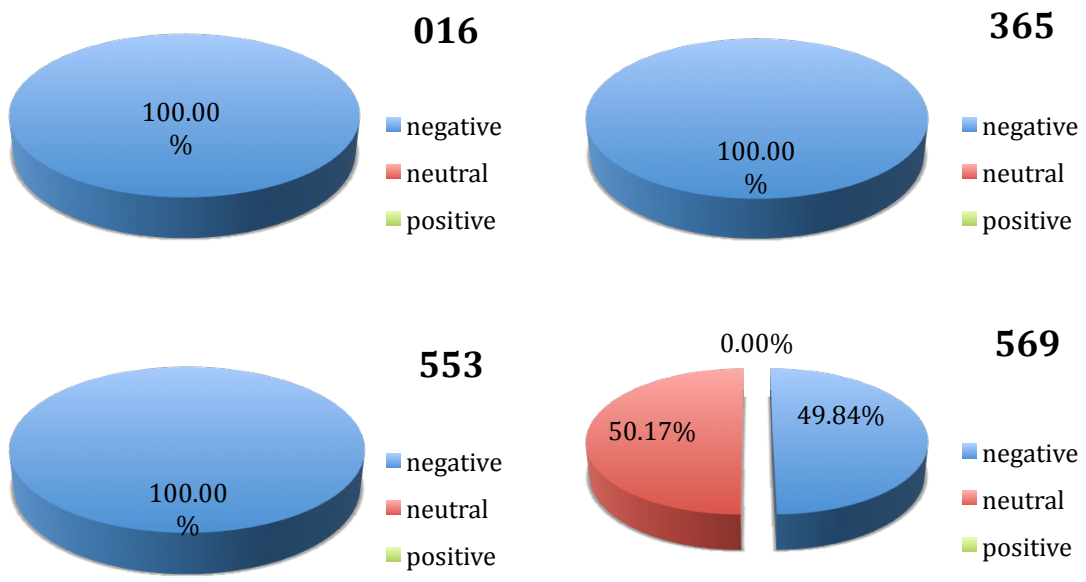


Figure 4:7 - Top left: Patient 016 (Case); Top right: Patient 365 (control); Bottom left: Patient 553 (control); Bottom right: Patient 569 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Group 2

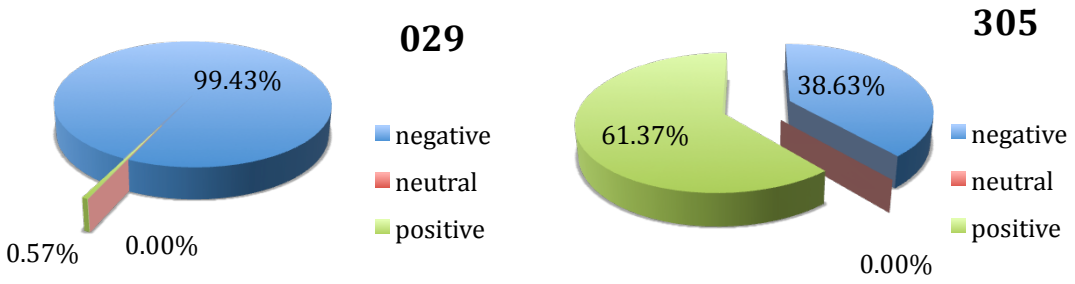


Figure 4:8 - Left: Patient 029 (Case); Right: Patient 305 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Group 3

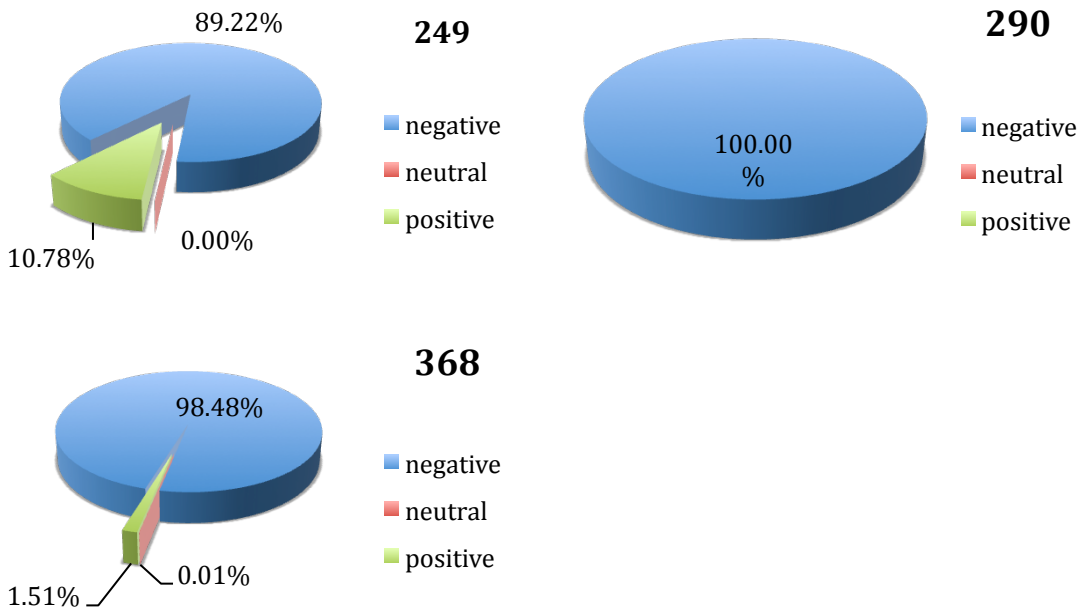


Figure 4:9 - Top left: Patient 249 (Case); Top right: Patient 290 (control); Bottom left: Patient 368 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Group 4

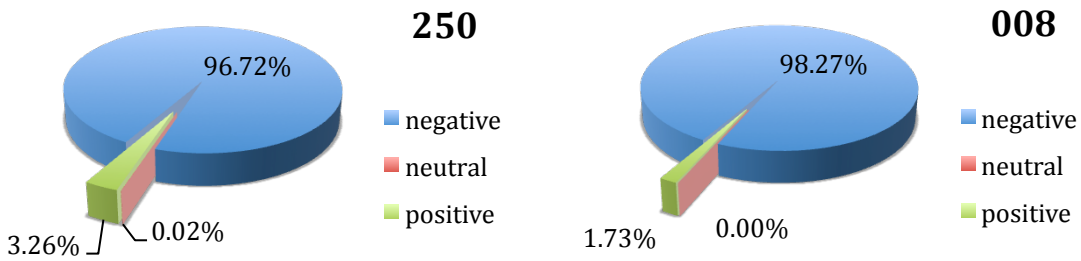


Figure 4:10 - Left: Patient 250 (Case); Right: Patient 008 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Group 5

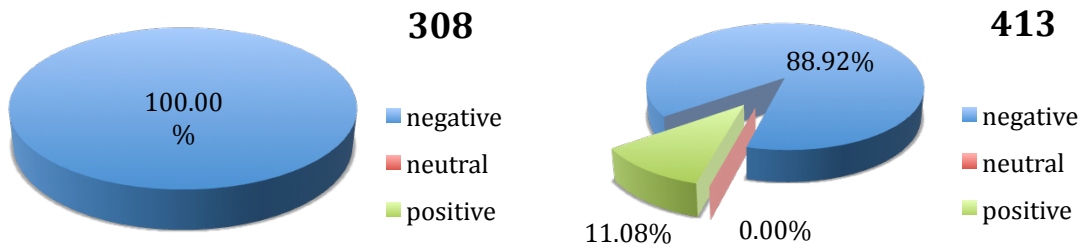


Figure 4:11 - Left: Patient 308 (Case); Right: Patient 413 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Group 6

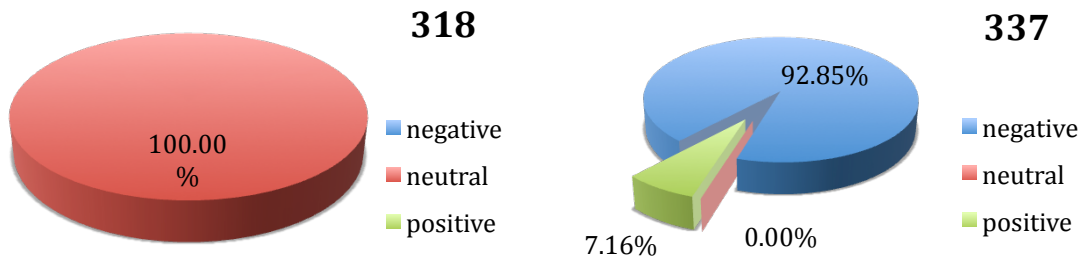


Figure 4:12 - Left: Patient 318 (Case); Right: Patient 337 (control). Charts show the proportion of sites under negative, neutral, and positive selection.

Table 4:10 - PAML Model M2a results. This table shows the proportion (p) of sites under negative, neutral, and positive selection, and the corresponding omega (ω) values. *White – Cases; Blue – Controls. P – proportion; ω = omega.*

Patient		Negative	Neutral	Positive
016	*p*	1	0	0
	ω	0	1	1
365	*p*	1	0	0
	ω	0.10884	1	1
553	*p*	1	0	0
	ω	0.45519	1	1
569	*p*	0.49835	0.50165	0
	ω	0	1	1

Patient		Negative	Neutral	Positive
029	*p*	0.99432	0	0.00568
	ω	0	1	513.17807
305	*p*	0.38626	0	0.61374
	ω	0	1	3.08971

Patient		Negative	Neutral	Positive
249	*p*	0.89222	0	0.10778
	ω	0	1	5.53348
290	*p*	1	0	0
	ω	0.06479	1	1
368	*p*	0.98478	0.00008	0.01513
	ω	0.65832	1	16.40407

Patient		Negative	Neutral	Positive
250	*p*	0.96719	0.00017	0.03264
	ω	0.06358	1	17.75726
008	*p*	0.98274	0	0.01726
	ω	0.20074	1	21.3079

Patient		Negative	Neutral	Positive
308	*p*	1	0	0
	ω	0	1	1
413	*p*	0.88923	0	0.11077
	ω	0	1	14.40761

Patient		Negative	Neutral	Positive
318	*p*	0	1	0
	ω	0	1	4.58636
337	*p*	0.92845	0	0.07155
	ω	0	1	29.46123

These results imply that there is a high level of negative selection occurring both in active and inactive disease patients, with seven patients demonstrating 99-100% negative selection. Of the subjects who demonstrate positive selection, only subjects 249, 305 and 413 demonstrate omega values within the accepted range ($0 < \omega < 15$). Consistent with this result, these patients also demonstrate the highest levels of diversity and amino acid changes in other analyses. Several patients demonstrate no sites under positive selection on internal branches.

A signed rank test was performed to elucidate if there was a difference in the proportion of sites under positive selection between active and inactive disease subjects. The data are shown in Table 4:11 below. A non-significant p-value of 0.44 was calculated.

Table 4:11 - Signed rank test of the proportion of sites under positive selection (p=0.44)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0%	0%	0%	0%	0%	0%
2	0.57%	61.37%			61.37%	60.8%
3	10.78%	0%	1.51%		0.755%	-10.025%
4	3.26%	1.73%			1.73%	-1.53%
5	0%	11.08%			11.08%	11.08%
6	0%	7.16%			7.16%	7.16%

PAML analyses and study power:

Subjects who demonstrate omega values above the level of acceptance must be treated with caution. It is likely that this is an artefact caused by a very small number of sites (e.g. – one or two) experiencing a non-synonymous change, whilst the remainder of the genome demonstrated no changes. For example, case subject 029 (group 2) demonstrates one non-synonymous change (See Table 3:11, p50), and an omega value of 513. The low diversity seen in the core gene of all subjects is thought to have influenced this result.

Results obtained from model M2a must be validated by comparison to model M1a. Model 1a only allows for two classes of selection (negative and neutral), thus the comparison of the two (using a Chi squared test, the "*likelihood ratio test*"), determines if the data best fit into one model.
In the current analysis the LRT resulted in only one of fifteen subjects exhibiting a statistically significant difference between models M2a and M1a, indicating that there is insufficient phylogenetic data to strongly support model M2a over M1a for every subject.

In summary, this analysis shows that greater power is required for PAML to accurately elucidate the proportion of sites that are under positive selection. The extraordinarily high omega values provide a valid explanation for the discrepancy between this result and the significant result obtained from the internal branch length analysis (Table 4:9, p81)

The table of the LRT results can be found in the appendix.

3. Sites under positive selection within patients

The Bayes Empirical Bayes analysis employed by PAML was used to elucidate sites under positive selection (Pr ≥ 0.50) within each patient individually. The data are shown in Table 4:12 below. Shading is used to differentiate between sites that are only found in cases or controls, or both.

Table 4:12 – Sites highlighted by PAML analysis model M2a. Red sites indicate a p-value < 0.05. *Green* – Positively selected in cases but not controls; *White* – Positively selected in both cases and controls; *Blue* – Positively selected in controls but not cases.
*** = p≤0.10; ** = 0.1<p≤0.15; * = 0.15<p≤0.2

	Active disease subjects						*Inactive disease subjects*								
Site	016	029	249	250	308	318	008	290	305	337	365	368	413	553	569
12													✓ **		
13			✓ **												
21		✓ ***							✓ **				✓		
22													✓		
24									✓						
27									✓						
28									✓						
36													✓		
48						✓			✓						
59									✓						
67			✓ *							✓ **					
74									✓						
80													✓		
91			✓ *						✓						
94									✓						
97									✓						
100									✓						
105									✓						
108									✓						
109							✓ ***								
113										✓			✓ **		
116										✓					
130			✓ *									✓			
149									✓						
151				✓											
156													✓		
165										✓					
168						✓									
181									✓						
182							✓ **			✓					

A signed rank test was performed on the number of positively selected sites (Pr ≥ 0.50), and a non-significant p-value of 0.31 was calculated (see Table 4:13 below). This indicates that there is no significant difference between cases and controls in relation to the number of sites under positive selection.

Table 4:13 - Number of positively selected sites, Signed Rank test (p=0.31)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	0	0	0	0	0
2	1	15			15	14
3	4	0	1		0.5	-3.5
4	1	2			2	1
5	0	7			7	7
6	2	5			5	3

Number of non-synonymous changes within patients

MacClade was used to calculate the number of non-synonymous changes that occurred on internal branches. A signed rank test was performed, and a non-significant p-value of 0.34 was calculated. The data are shown in Table 4:14.

Table 4:14 - Signed rank test of the number of non-synonymous changes occurring on internal branches, (p=0.34)

	Case	Control 1	Control 2	Control 3	Control Mean	Δ
1	0	1	1	1	1	1
2	1	2	4		3	2
3	8	1	7		4	-4
4	3	5			5	2
5	0	8			8	8
6	2	3			3	1

4.3.2.3 - Summary of parameters used to compare cases and controls

In this section (4.3.2), we examined the six amino acids under positive selection pressure, and demonstrated an almost significant difference in the mutation of these sites in active and inactive disease (p=0.06).

In addition, we separately examined the between-patient and within-patient selection pressure, since it is possible that two different mechanisms are at play. We found no significant difference (p=1.0) between active and inactive disease in external branch lengths (between-patient). An examination of the number of non-synonymous changes on external branches was also non-significant (p=1.0). An examination of the processes occurring within-patients (internal branches) demonstrated a statistical significance when internal branch lengths were compared (p=0.03). This finding was not ratified by further analyses, however a trend towards less selection pressure in active disease subjects is observed, especially if subject 249 is considered an outlier (as discussed in section 3.4.5, p61). An examination of the data from the within-patient analyses (Table 4:15, section 3) corroborates this fact.

Table 4:15 - Summary of parameters used to compare cases and controls, showing the p-values from the signed rank tests

Analysis	Case	Control	p-value
Non-synonymous changes at 6 amino acids	11	18.5	0.06

Analysis	Case	Control	p-value
External branch lengths	0.136	0.153	1.00
Non-synonymous changes on external branches	21	22.2	1.00

Analysis	Case	Control	p-value
Internal branch lengths	0.024	0.073	0.03
Non-synonymous changes on internal branches	14	24	0.34
Proportion of sites under positive selection	14.61	82.095	0.44
Number of positively selected sites	8	29.5	0.31

4.4 Discussion

In this chapter, we investigated selection pressure in two ways. Firstly, by conducting an analysis of the selection pressure occurring on the whole tree, and secondly by using six paired groups to analyse differences between *within-patient* and *between-patient* selection pressure. We conclude this chapter by discussing the mechanisms involved in between-patient and within-patient selection.

4.4.1 Whole tree analyses

The branch-site model is designed to test whether certain branches exhibit differential selection compared to the rest of the tree. We used this model to compare the level of selection occurring within inactive disease patients, against a 'background' of active disease patients (and vice versa). Consistent with our hypothesis, this analysis found approximately 3% of sites that were differentially selected between inactive and active disease ($\omega = 7.74$) (Table 4:1, p69).
A comparison in which active disease clones were labelled as the foreground, and eAg+ and eAg- inactive disease clones were labelled as the background, demonstrated a distinct lack of positive selection within active disease clones. As Table 4:1 demonstrates, 0% of sites within active disease patients exhibit any form of positive selection ($\omega = 1$).

A Bayes Empirical Bayes analysis then elucidated six sites that were under positive selection ($Pr \geq 0.95$) in the whole tree. This was reinforced by an analysis of the Most Ancestral Clone, and is also consistent with previously described results[45].

A collective analysis (such as the branch-sites model) compares *all* cases or controls to the background, eliminating the problem of only identifying sites under positive selection that were specific to one patient. However this does not eliminate the problem presented when both the foreground and the background share the *same* sites under positive selection. An interesting future study would be to separately compare controls and cases, to the pre-seroconvertant (eAg+) clones. This would highlight sites that have changed during seroconversion.

Furthermore, it is possible that the strong omega value seen in this analysis is a result of different numbers of case and control patients (n=6, n=10; respectively). Therefore, further analyses were required to confirm these findings. A paired study was employed, using six HLA class-I matched groups.

4.4.2 Comparisons of selection pressure in HLA class-I matched groups

Initially, we conducted an analysis of the number of patients demonstrating non-synonymous mutations found at the aforementioned six amino acids. This revealed an almost significant p-value of 0.06. This then prompted further investigation into the subtypes of selection pressure (fixed and unfixed) occurring, to see if either of these gave a greater indication of the difference between the groups, as well as some clues to the mechanisms involved.

Between-patient analyses

The examination of the external branches (fixed mutations) showed no significant difference between the groups in external branch lengths (p=1.0), nor in the number of non-synonymous mutations occurring on external branches (p=1.0). This indicates that the process that drives fixation, functions indiscriminately between active and inactive disease. It is unclear whether transmission selection, or immune-mediated purifying selection, plays a dominant role in determining external branch length.

In addition, the Most Ancestral Clone analysis highlighted 16 amino acids that were found to be under some form of positive selection (Pr ≥ 0.50). Four of these sites (77, 113, 130, and 180) were also found in the analysis of the whole tree. The remaining sites are thought to be under negative selection within patients, therefore are not highlighted by PAML.

Within-patient analyses

The analyses conducted on within-patient (unfixed) mutations demonstrated a statistically significant (p=0.03) difference in internal branch length, indicating that the immune processes occurring *within patients* are significantly different between active and inactive disease.
We therefore conducted an analysis of the selection pressure occurring on internal branches. Unfortunately, there was not enough power in the selection analyses to determine whether this significant finding was a positive selection effect, although the data do not exclude this possibility.

The Bayes Empirical Bayes analysis employed by PAML (Table 4:12, p88) elucidated six sites under weak selection pressure ($Pr \geq 0.50$) and two sites under strong selection pressure ($Pr \geq 0.95$) within active disease clones, as well as 24 sites under weak selection pressure and six sites under strong selection pressure within inactive disease clones. In addition, only three sites were found that are specific to case subjects. 22 sites specific to control subjects were found.
The signed rank test performed on this data failed however to give a significant result (p=0.34).

Interestingly however, the sites that were highlighted correlate well with results from other analyses[45], indicating perhaps that the lack of significance is a product of small sample size rather than lack of positive selection.

Comparison of Between and Within-patient selection

Between-patient and within-patient selection was measured separately since it was hypothesised that different selection processes could be occurring in each. Section 4.3.2.2A (p74) demonstrated that we were unable to elucidate any significant differences occurring between patients, between active and inactive disease. This is thought to indicate that the process acting on the external branches occurs indiscriminately in both case and control. One possible hypothesis is that external branches represent transmission selection, where the act of transmission places selection pressure on the virus. Currently, there is no known difference in the method of transmission between active and inactive disease clones. Therefore, similar selection pressure could be expected on external branches. Alternatively, external branches may represent the effects of purifying selection during HBeAg seroconversion, leading to the fixation of a particular amino acid. Since both active and inactive disease clones have undergone seroconversion, similar selection pressures would be expected on the external branches. Interestingly, an examination of the level of fixation at known hotspots demonstrated a greater number in active disease clones (Appendix Table 8:12, p160)

Section 4.3.2.2B (p81) demonstrated a significant difference ($p=0.03$) between internal branch lengths within patients, however this finding could not be ratified by selection analyses, although this is likely a result of insufficient power, as shown by the extra-ordinarily high omega values.

The internal branches are representative of changes that have occurred within the patient. However it is unknown as to whether all clones within a clade are descendent from a single ancestor, or are in fact a result of multiple infections. It is assumed to be former, given the (generally) low intra-patient diversity seen in the phylograms.

In light of these differences, a comparison of the sites highlighted by the Bayes Empirical Bayes analyses demonstrated seven sites (see Table 4:20, below) that are differentially selected between internal and external branches, supporting the hypothesis that different processes are involved internally and externally.

Table 4:20 – Comparison of sites under positive selection between patients, with sites under selection within patients. Ticks represent a site that is identified with > 50% probability as being under positive selection.

Site	Between Patients	Within Patients
13	✓	✓
26	✓	
49	✓	
64	✓	
79	✓	
83	✓	
91	✓	✓
109	✓	✓
113	✓	✓
130	✓	✓
149	✓	✓
151	✓	✓
177	✓	
180	✓	

The issue of Between-patient and Within-patient selection pressure remains controversial, since the exact selection mechanisms that are acting on HBV during its evolution are still somewhat uncertain. It would be errant to assume that all changes on a branch (external or internal) were due to a single force. Theoretically, it is possible that these mutations have resulted from transmission, or from strong selection pressure resulting in fixation, or both. In reality, the number of transmission events that a branch represents is impossible to calculate. Additionally, the act of transmission itself could exert a selection pressure, thus selecting for a certain genetic element. Therefore, the changes that are observed *between*-patients are of interest, but are not *necessarily* immune mediated.

Unlike HIV, HBV is not known to demonstrate transmission chains, since the eAg is required to establish a chronic infection[1, 9, 23]. All chronically infected subjects are therefore thought to have been infected by an HBeAg positive person, thus we believe that transmission is not an important contributor to long external branches. This assumes however that all seroconvertants (HBeAg negative) do not possess any eAg +ve clones. Chapter 5 of our study addresses this assumption.

Further considerations

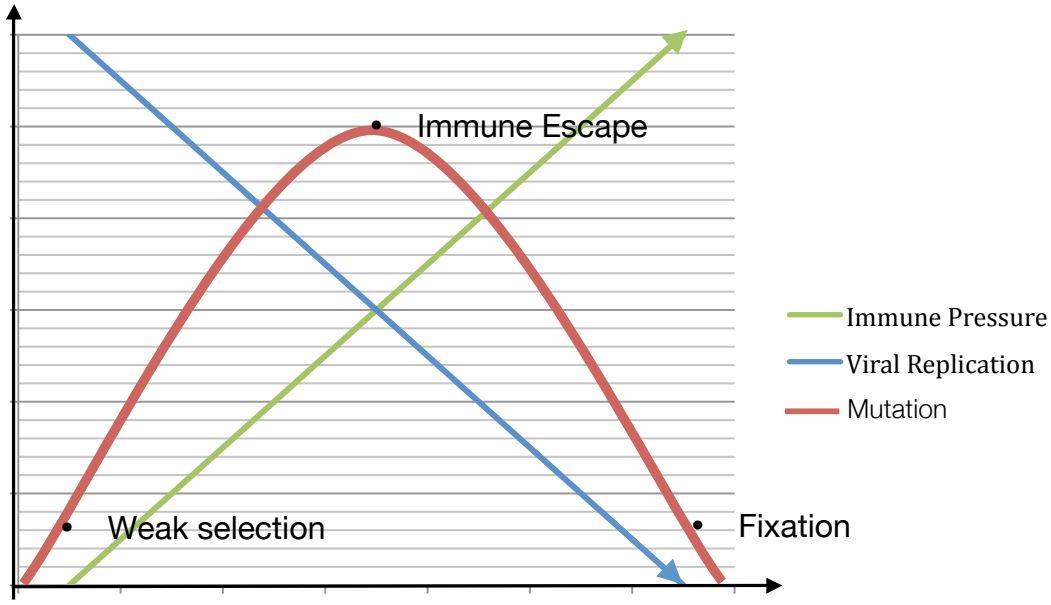


Figure 4:13 - Schematic of possible mechanism in HBV evolution

The figure above summarises a possible mechanism occurring within HBV evolution. The green arrow represents increasing immune selection pressure being placed on the virus, following HBeAg seroconversion. The blue arrow represents the decreasing level of viral replication, which is inversely proportional to immune pressure. The red curve however represents the mutational ability of HBV. Pre-seroconversion, viral replication is high, and conversely immune pressure is low. At this point the level of mutation is low (due to a lack of selection pressure). Following seroconversion, immune pressure increases, causing an inversely proportional decrease in viral replication levels. A critical point is reached at the intersection of the two however, as this represents the optimal time for immune escape mutations to arise in the virus. This is because the levels of immune pressure are sufficiently high to generate selection pressure, which is then propagated by a moderate level of viral replication. Therefore, if the immune pressure does not progress past this point, either due to an inherent deficit or suppression, this state will persist.

4.4.3 Summary

Therefore, overall, we could not detect a large difference in selection pressure between the six HLA class-I matched groups, however a subtle defect is not excluded. Examination of Table 4:15 (p90) and Table 3:12 (p55) demonstrates a consistent trend towards less diversity and selection pressure in active disease subjects.

A possible mechanism for this is thought to be a combination of the tolerising effects of the eAg and frequency dependent selection. This is developed in chapter 5.

-

Chapter 5:

The Effect of the E Antigen

5 The effect of the e Antigen

5.1 Introduction

Recent evidence suggests that the E Antigen (eAg) is purported to cause immune suppression, most especially until the time of seroconversion. Whilst this has not been conclusively proven, the evidence is compelling[6]. Due to the purported suppressive effect of the eAg, we hypothesised that this may play an integral role in the viral infection and replication dynamics, specifically in disease persistence. Given that our original study was focused only on eAg- clones (post-seroconvertant patients), examination of the eAg+ clones (pre-seroconvertant patients) fell outside of the scope of our project. Nevertheless, we felt that the data compiled in our original study was consistent with the new hypothesis, sufficiently enough to warrant discussion in this chapter.

5.2 Frequency Dependent Selection

Frequency Dependent Selection (FDS), also known as Rare Allele Advantage, is likely to play a significant role in the viral dynamics of HBV within a host.
FDS is when the *selection* pressure acting on a trait or an individual is *dependent* on its *frequency* in the population.
John von Neumann and Oskar Morgenstern developed 'Game theory' in 1944 in the book "Theory of Games and Economic Behaviour"[46], and is an attempt to model, mathematically, the behaviour of individuals based on the behaviour of other individuals. Game theory has been applied to politics, economics, philosophy, psychology and biology, and has been shown to fit to biological data with great accuracy.
When considering *fitness*, (that is, the evolutionary ability of an organism to survive in a given environment), Game theory can be applied as rare allele advantage / frequency dependent selection.
In the case of HBV, frequency dependent selection may work according to the following four stages.

Stage 1

In the early stages of infection, all clones are eAg+. Whilst metabolically costly, the survival advantage to the virus is great, given the immunosuppressive effects of the eAg. This is reflected in the high levels of immune suppression and low levels of selection pressure (as demonstrated in Figure 5:2, p102) Over time, some viruses may lose their eAg, either due to random mutation or perhaps immune selection pressure. At an individual level, this would be a deleterious mutation, and this virus would be removed via the actions of negative selection. However, in a collective environment (i.e. > 1 virus), this single eAg- virus can "cheat" by relying on the immunosuppressive effects of its eAg+ siblings, thus continue to survive and replicate, creating further eAg- viruses. The eAg clone is also metabolically advantaged, as it does not have to produce the eAg. In this way, the overall advantage lies with the eAg- clones. In Figure 5:1, this is represented by the 'balance of advantage' being weighted in favour of the few eAg clones.

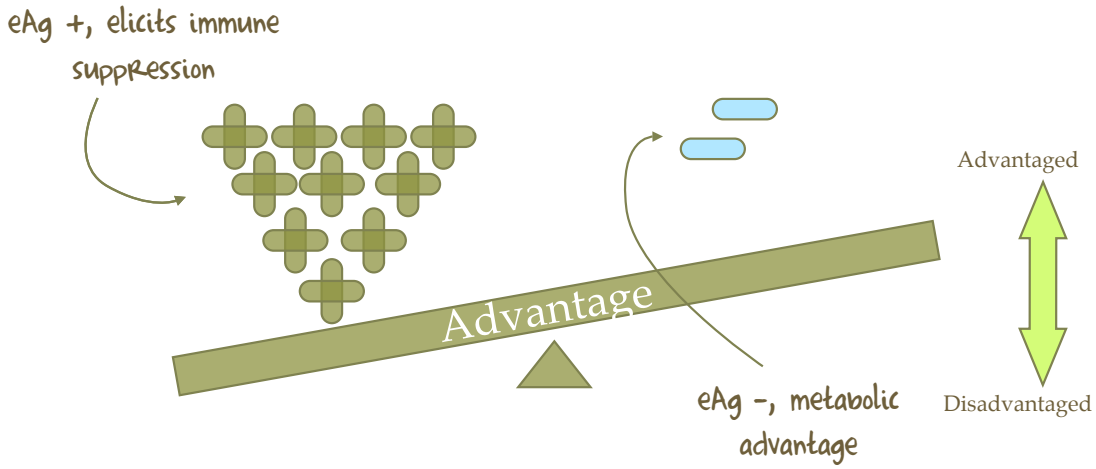


Figure 5:1 - Frequency Dependent Selection. Stage 1 – eAg+ » eAg-

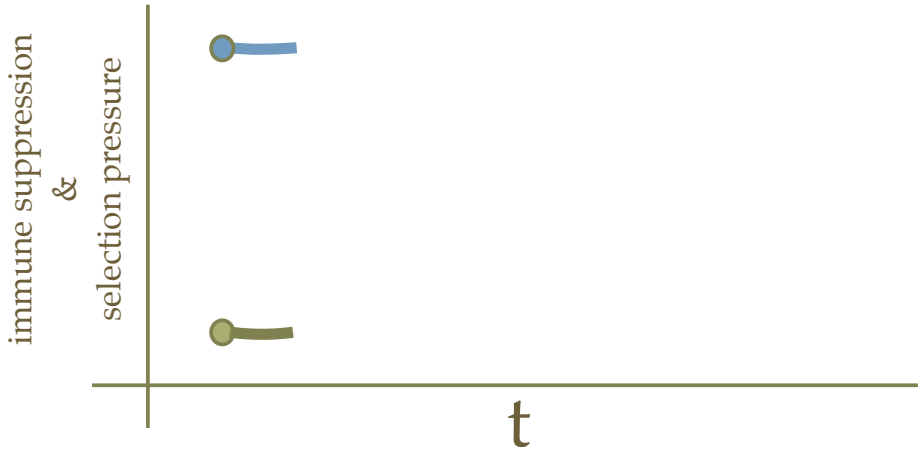


Figure 5:2 - Immune selection pressure (green) and immune suppression (blue). Stage 1 - eAg+ » eAg-

Stage 2

Over time, the reduction in the frequency of eAg+ clones by mutation and the simultaneous increase in the frequency of eAg- clones by mutation and replication (see Figure 5:3) results in a decrease in the level of immune suppression. This is illustrated by the blue line in Figure 5:4. A reduction in immune suppression conversely leads to an increase in selection pressure placed upon the virus, illustrated by the green line in Figure 5:4. In this way, the *selection* placed upon the viral clone (in this case, the eAg-) is *dependent* on its *frequency* in the population.

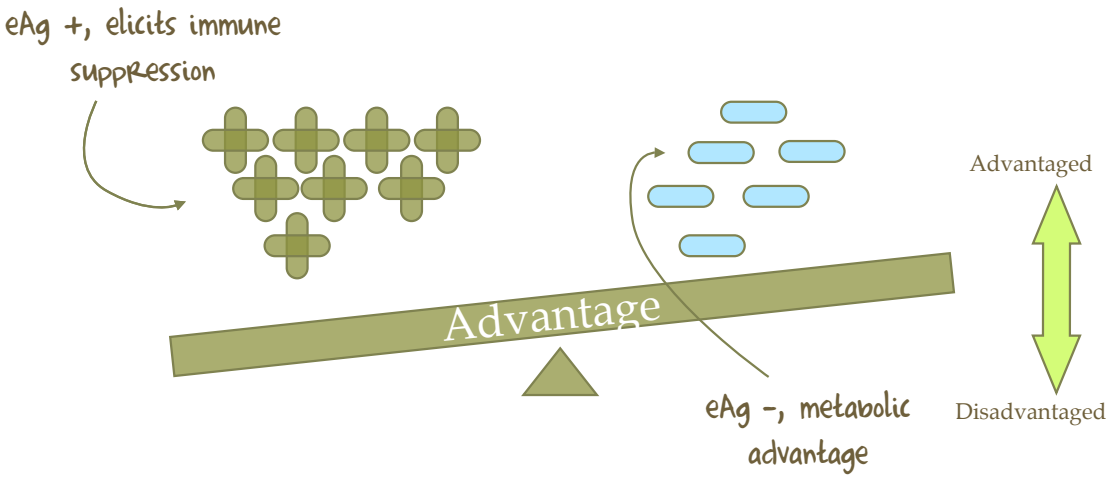


Figure 5:3- Frequency Dependent Selection. Stage 2 - eAg+ > eAg-

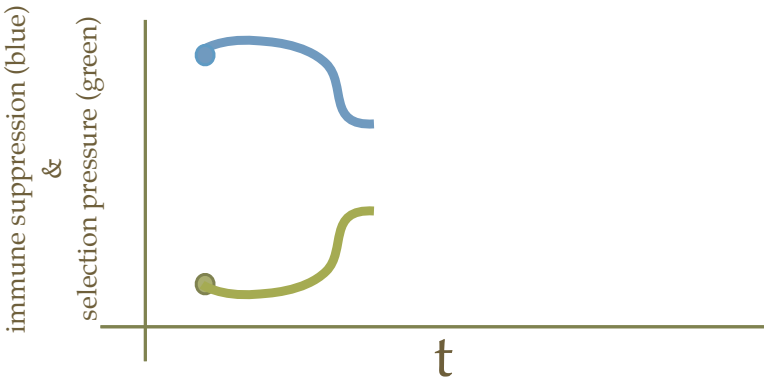


Figure 5:4 - Immune selection pressure (green) and immune suppression (blue). Stage 2 - eAg+ > eAg-

Stage 3

Ultimately, continuation of this process will result in equilibrium between the eAg+ and the eAg- virus population (see Figure 5:5). At this point, the advantage previously afforded to the eAg- population now becomes a liability. The level of immune suppression elicited by the 'pool' of eAg+ clones is no longer sufficient to cover the whole virus population, and the eAg- clones can no longer 'cheat'. As the frequency of the eAg- clone increases in the population, so does the selection that it experiences. As demonstrated by Figure 5:6, the level of immune suppression and selection pressure are also at a point of equilibrium.

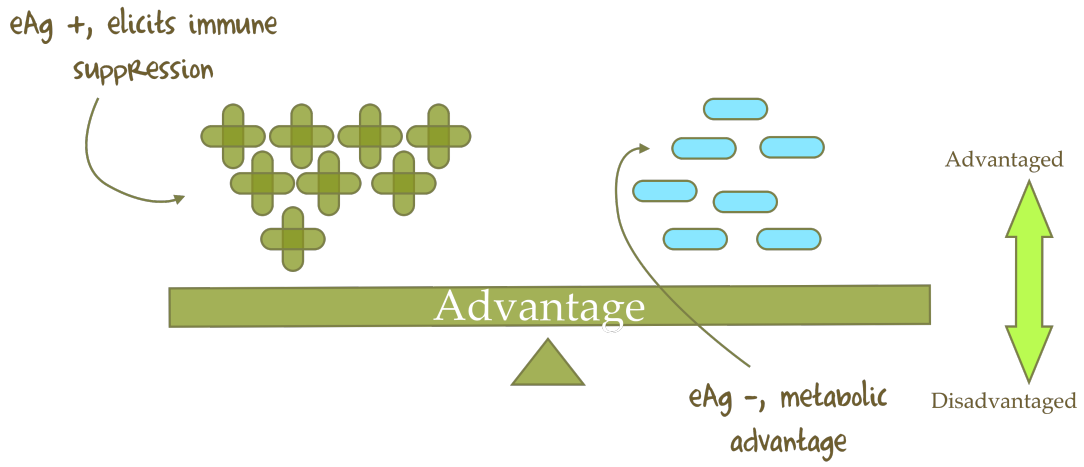


Figure 5:5 - Frequency Dependent Selection. Stage 3 - eAg+ = eAg-

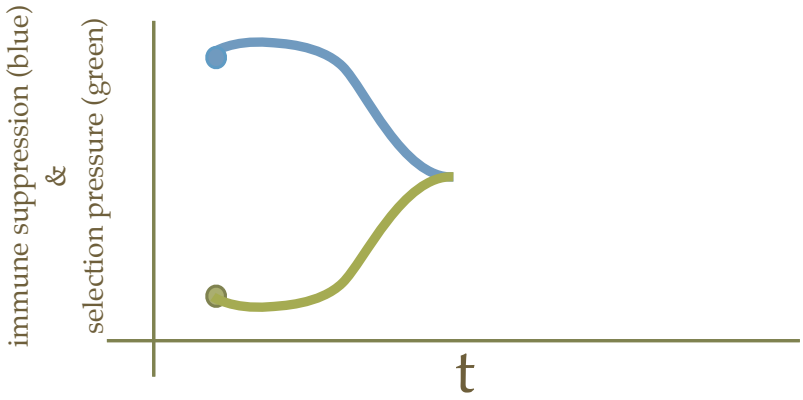


Figure 5:6 - immune suppression (blue) and immune selection pressure (green). Stage 3 - eAg+ = eAg-

From this point, there are two foreseeable options:

Stage 4a - Normal Immune system

If the patient has a robust immune system, it will hypothetically continue to elicit selection pressure on the virus, leading to a continual increase in eAg- clones. This could be thought of as the immune system getting “the upper hand”, evidenced through the appearance of stop codons in the eAg, and eventual ‘control’ of the virus. In this ‘unprotected state’, it would be expected that there would be a higher incidence of mutations in the eAg- clones, as a result of the immune pressure.

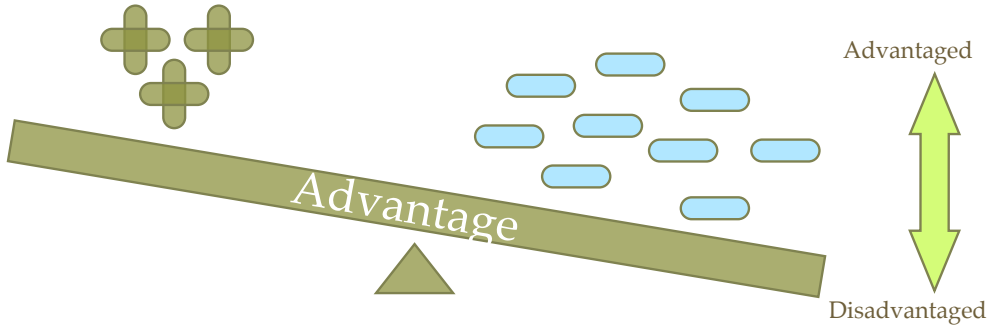


Figure 5:7 - "The Upper Hand" - eAg+ ≪ eAg-

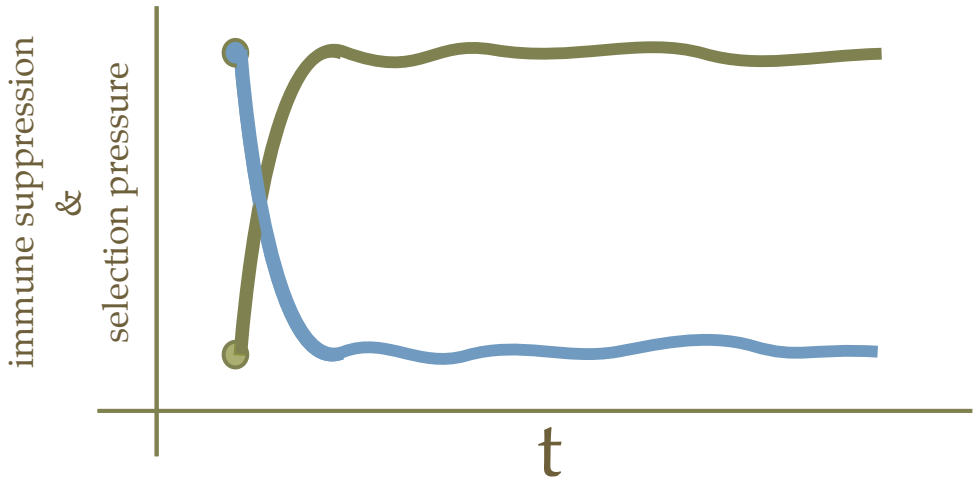


Figure 5:8 - Immune selection pressure (green) and Immune suppression (blue). "The Upper Hand": The immune system has effectively controlled the virus, with all clones eventually being eAg-

Stage 4b - Deficient immune system

If the patient has an immune deficit, then the 'strength' of the immune system to gain control is diminished. Up until the point of equilibrium, selective fitness rested with the eAg- clones, given their metabolic advantage and cheating strategy. However, at the point of equilibrium, selection will now favour eAg+ clones, given their immunosuppressive ability. This will result in the balance in the figure shifting to the left.

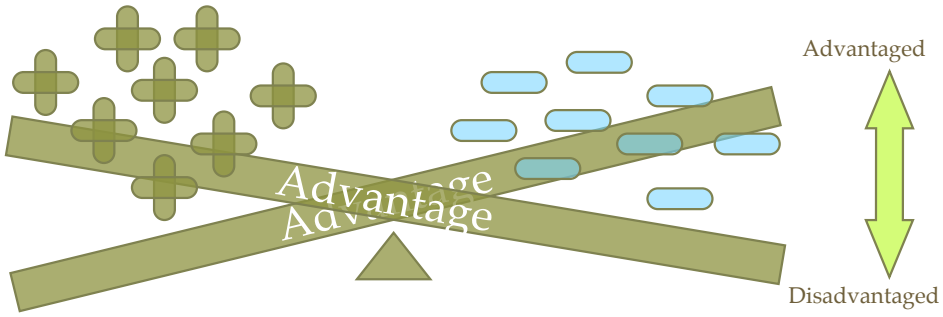


Figure 5:9 - A dynamic model, allowed to persist due to a deficient immune system

Without the pressure of a strong immune system, this process will ultimately result in a dynamic state of equilibrium, as the process goes back and forth. In this dynamic state, the amount of time that a clone is 'exposed' to the immune system is reduced. Thus it would be expected that eAg- clones in this category would evidence less mutations as a result of immune pressure.

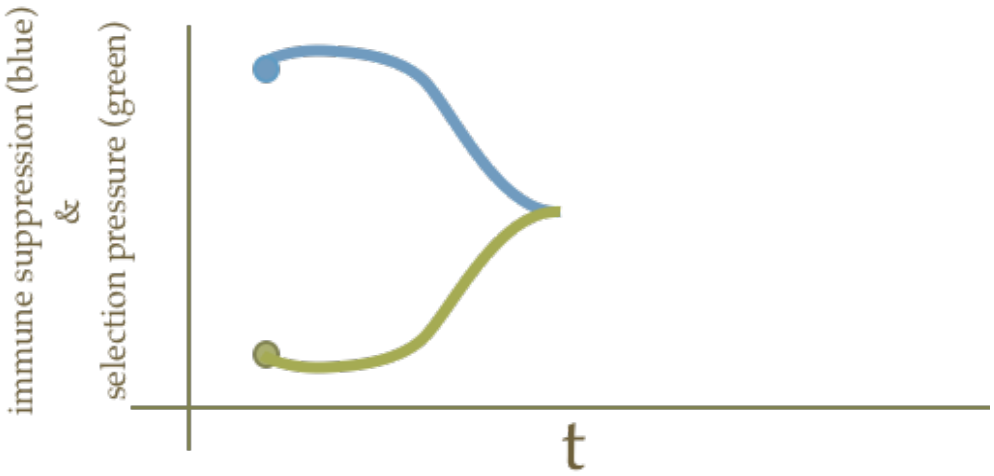
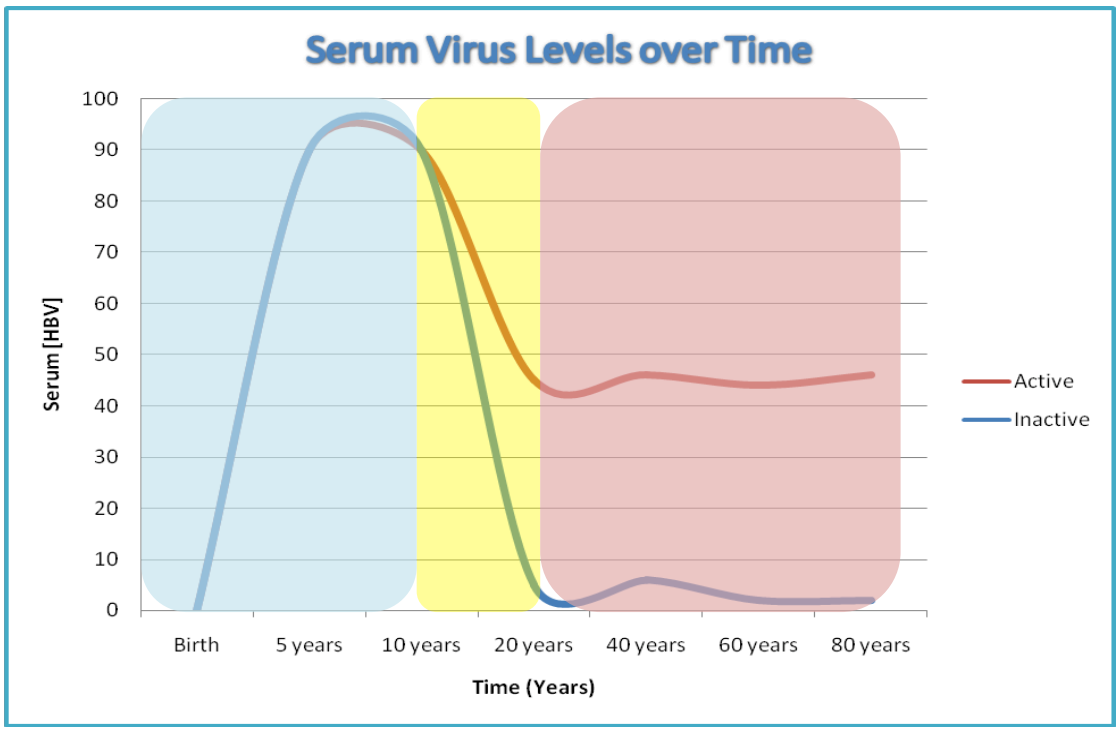


Figure 5:10 - Immune suppression (blue) and Selection pressure (green) in a dynamic model.

This hypothesis is consistent with the observed data. When Figure 1:1 (included again below) is re-examined with this hypothesis in mind, possibility exists that the active disease population are in the 'dynamic model', whilst in the inactive disease population, the immune system has 'gained the upper hand'. This may offer an attractive hypothesis to explain why the virus can persist within the host for an extensive period of time.



5.3 Sero-status versus geno-status

For the above hypothesis to be correct, this would require the presence of eAg+ clones within the post-seroconvertant CHB population (active disease).

An interesting discovery was made when analysing the phylogram of the core gene with the eAg region included. Inclusion of this region elicited moderate changes in the phylogram. Most notably, the location of clone 368F changed dramatically, resulting in a lack of clustering with the other clones extracted from patient 368. Re-examination of the alignment showed that despite being classified as a post-seroconvertant (eAg-), patient 368 still contained one viral clone (=7% of clones extracted from 368) with an intact e antigen ORF (see Figure 5:11)

Typically, upon seroconversion, loss of the e antigen occurs via the appearance of a stop codon immediately prior to the start codon for the core gene (G1896A), and yet as Figure 5:11 demonstrates, clone 368F does not contain the stop codon.

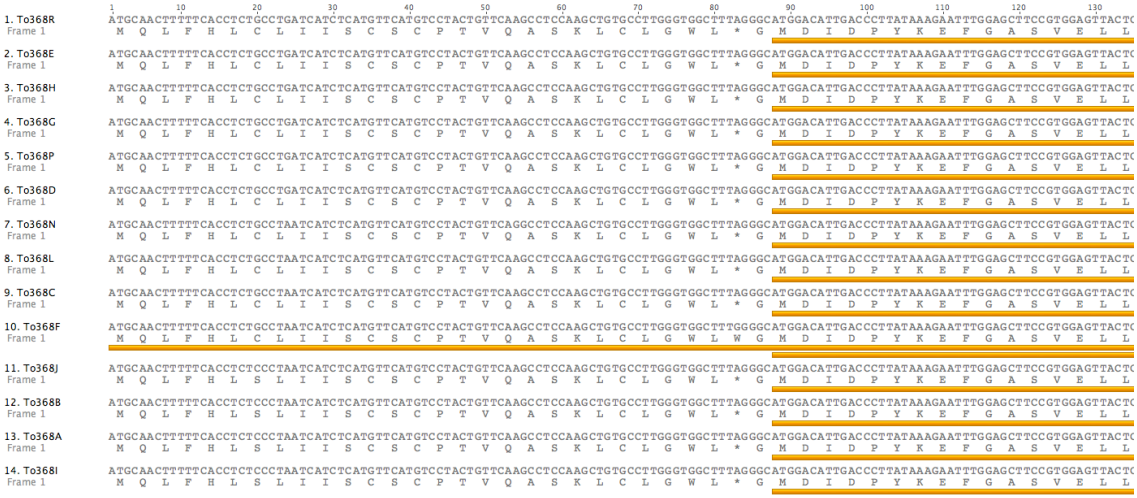


Figure 5:11 - Graphic display of intact eAg ORF in a serologically classified eAg negative patient (368).

This discovery prompted a search of the entire dataset to elucidate if there were other “post-seroconversion” patients that contained eAg+ clones, and conversely, if there were any “pre-seroconversion” patients with eAg- clones. Patients 460, 233, 544 and 391 all demonstrated this phenomenon. As can be seen in Figure 5:12 (p109), it is clear that the ‘sero-status’ of a patient does not refer to the absolute genotypic status.

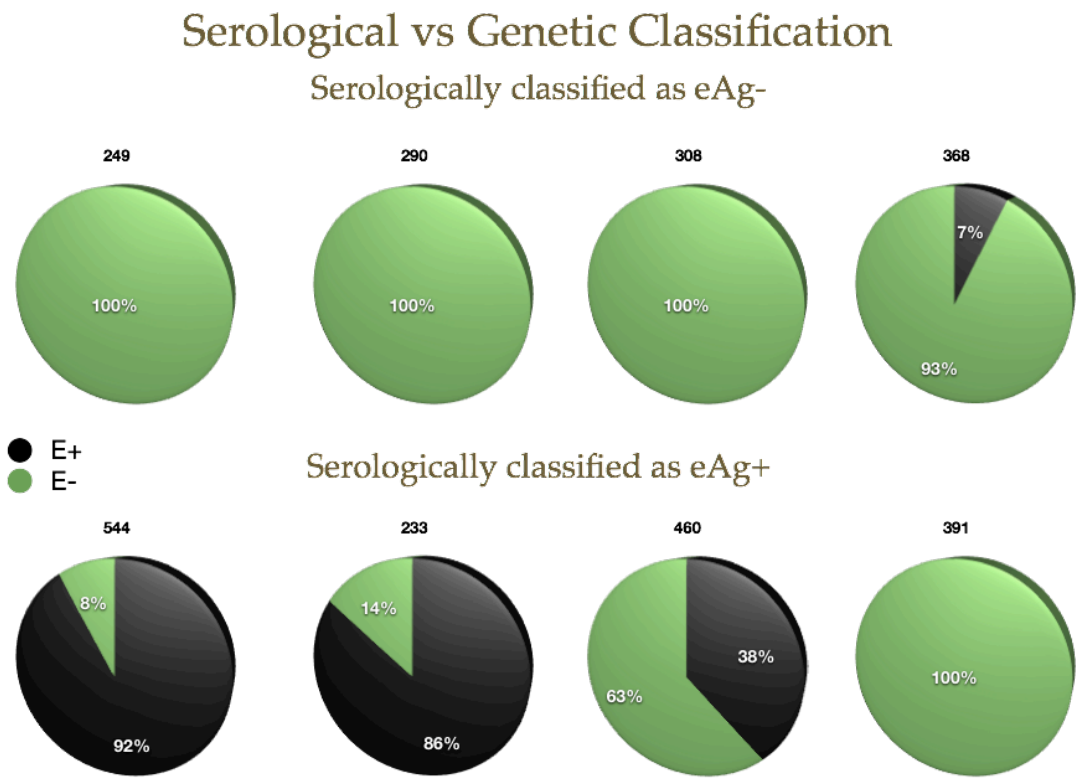


Figure 5:12 - Proportion of clones within selected patients who are eAg+ or eAg-

Unfortunately, patient 368 is not an active disease patient. However, since this fell outside the scope of our original project, limited data was gathered. Thus, further investigation may elucidate more eAg+ clones within CHB post-seroconvertant patients.

Chapter 6:

Summary and Conclusions

6 Summary and Conclusions:

6.1 Summary

HBeAg negative Chronic Hepatitis B is a major health concern worldwide, with current treatments being largely ineffective, and no cure available. Whilst a vaccine is available, its cost is preventative for developing countries where HBV in endemic. Furthermore, the vaccine is only prophylactic, thus offering no hope to those already affected. This study aimed to give meaningful clues into amino acid sites that could be incorporated into a therapeutic vaccine.

To achieve this goal, a greater understanding of the difference between active disease and inactive disease subjects is required. Recent evidence has suggested that a host CD8 T cell deficit may be responsible for this difference. Direct immunological assays have not elucidated any significant difference, and a new approach is needed. In HBV, the core gene is thought to be a major target of the immune response. We therefore designed our study to test for selection pressure differences in the core gene between HBV clones extracted from active and inactive disease subjects, using phylogenetic and computational methods.

Our study design matched six active disease subjects with 10 inactive disease subjects based on their HLA class-I alleles. This gave us sufficient power to detect a large selection pressure deficit between the groups, based on previous studies. We did not expect to elucidate a subtle difference. To achieve this, we extracted, cloned and sequenced the HBV core gene from each subject, which were then aligned using Clustal-W and subjected to phylogenetic and computational analyses, using PAML, PAUP*, PhyML and MacClade.

In chapter 3, we aligned the cloned sequences and removed all nonsense mutations. We constructed a maximum likelihood phylogram of the core gene, and confirmed its validity based on previously described results, and using the SH test. We then examined the diversity within the data, and demonstrated a consistent trend towards decreased mutations in the active disease population.

In chapter 4 we examined the selection pressure that was occurring on the whole tree, the external branches, and the internal branches. We also examined the sites

that were highlighted as being under some form of positive selection.
On the whole tree, we demonstrated increased proportion of sites (~3%) under selection pressure in control clones, with an omega of 7.74. In addition, we elucidated six amino acid sites that were highlighted by PAML as being under positive selection in two separate analyses. It is however possible that the significant omega value observed is due to different sample sizes between active and inactive disease, prompting a further investigation into the within-patient and between-patient differences. An examination of the number of subjects who demonstrated a non-synonymous change at the six amino acids was almost significant (p=0.06), supporting the omega value, and indicating that this study design is worth continuing.
We had good power to detect differences in external branch lengths, and were unable to elucidate any large difference between active an inactive disease. However, it is thought that mutations due to transmission effects may be a confounding factor, as they will influence the sites that are under positive selection, however it cannot be determined if this is immune or transmission mediated.
In addition, we detected a significant difference (p=0.03) in internal branch lengths, however we had insufficient power to detect if this difference was related to positive selection pressure.

Furthermore, we presented a novel hypothesis explaining the mechanism for the lack of suppression of viral replication in chronic active disease, involving the immunosuppressive effects of the HBeAg and the documented phenomenon of frequency dependent selection.

In this study, we tested three major hypotheses.

1) There is a major deficit in selection pressure in chronic active disease patients
2) There is a limited selection of amino acids that are targets of immune activity
3) There is a difference in the sites that are targeted by the immune system

6.1.1 Hypothesis 1: There is a major deficit in selection pressure in chronic active disease patients

Our study clearly did not exclude this hypothesis, demonstrated by the strong omega value (7.74) and the corresponding six amino acids highlighted by two separate analyses as being under positive selection. Although no significant p-value was obtained in further analyses, our data demonstrate a consistent and strong trend towards a deficit in active disease patients. Table 3:12 (p55) and Table 4:15 (p90) display this trend, with all analyses resulting in a higher value in inactive disease subjects. In addition, greater significance could be attained if case subject 249 (group 3) is considered an outlier (as discussed in 3.4.5, p61).

It is thought therefore that further study should be conducted on the core gene using this study design, as significance was almost attained. However, the following limitations should be addressed.

Power

It is clear that a greater number of class matched subjects is required to achieve sufficient power in further analyses, and to mitigate any effects of outliers. It is predicted that a doubling of the sample size (case = 12, control = 20) would be sufficient, however this does present a problem as subject recruitment is difficult.

This could be addressed in three ways. Firstly, we could wait for a greater number of patients to present at the Langimalie clinic where the current study was centred. This may take some time. Secondly, the study could be expanded to include any subject of Polynesian ethnicity, who have the same viral genotype and HLA class I alleles. However, this will require further ethical consent requirements, and will be logistically challenging. In the current study, recruitment was achieved via the Langimalie Tongan Hepatitis clinic, thus providing a central location for Tongan subjects. A study that includes all Polynesians will require some method for centrally recruiting and sampling these patients. Thirdly, the study could be re-designed to focus on Chinese subjects, allowing a greater sample size. Chinese subjects however are known to have greater heterogeneity in their HLA class-I alleles[1], thus making the class matching protocol more difficult.

Cross-sectional

A further limitation in this study was the restriction in inferences that can be drawn from a cross-sectional study, which provides only a 'snap-shot' of viral evolution within the host. This could be addressed by modifying the study design to allow serial sampling of subjects, thereby creating a longitudinal study of mutation repertoires. This would allow a more accurate inference of evolution to be made, based on the mutational history of the viral sequence.

Other regions of the genome

Our study was limited to the core gene, and to some extent the pre-core region (eAg). While it is thought that the core gene represents a major immune target, a recent paper by Riedl *et al*[47] has demonstrated some preliminary success in elucidating epitopes specific to the surface/envelope gene.

Lack of diversity

Another challenge encountered was the lack of diversity within the dataset. Since phylogenetic analyses are largely based on diversity ("signal"), a lack of diversity confines the ability of the phylogeneticist to accurately elucidate evolutionary patterns, and limits the computational power of the phylogenetic software algorithms. Our analyses found that increasing the number of clones taken from one patient did not affect the diversity measure, therefore indicating that sample size is not the reason for lack of diversity.

6.1.2 Hypothesis 2: There is a limited selection of amino acid sites that are targets of immune activity

Our study highlighted six amino acids that are under strong positive selection in the whole tree, and ten sites that are under weak selection pressure. Therefore this hypothesis is not excluded. The six strongly selected sites are discussed here.

Strongly Selected sites (21, 26, 77, 113, 130, 180)

Site 21

Site 21 is of special interest, as it is found within a known HTL epitope (1 – 25) and the well documented CTL epitope (18 - 27). Site 21 demonstrated a peak for the number of substitutions (Figure 6:1, p123) with four changes at nucleotide 61 and three at nucleotide 62 in the lower histogram, and ten changes at nucleotide 61 and four at nucleotide 62 in the upper histogram, as well as demonstrating statistically significant ($p < 0.05$) positive selection in one patient in the PAML M2a analyses of internal selection. Site 21 also displayed non-synonymous changes in seven subjects.

Examination of the alignments revealed that all active disease patients contained a serine at residue 21, indicating strong fixation, with the exception of subject 029 who is genotype D4. This trend was not seen in the inactive disease patients. Therefore, site 21 indicates a significant site for further investigation given the above observations. A study with increased power would benefit from building on these findings.

Site 26

Site 26 was also found within the well documented 18 – 27 CD8 T-cell epitope, and was highlighted as being under positive selection ($p = 0.02$) on external branches. Within active disease, a fixation of Serine was seen in all clones, which is not repeated in inactive disease clones who demonstrated Serine, Proline and Asparagine. Site 26 was not highlighted by any within-patient analyses, and likely represents a site that is important in between-patient selection.

Site 77

The highlighting of site 77 is consistent with previous studies in Tongans[45] and demonstrates a p-value of 0.01 both in the whole tree, and in the external analysis. Both case and controls demonstrated non-synonymous changes at this residue.

Site 113

The analysis of external selection (p80) showed that site 113 demonstrated an almost significant p-value of 0.06, which infers that site 113 is important in external selection. Site 113 is within a major B-cell antigenic determinant region, found between residues 107-118, and demonstrates a total of nine changes in the upper histogram (Figure 6:1, p123), with a change at all three positions. The lower histogram shows six changes at site 113, internally. These are all found in controls. Interestingly, these six changes are all third position changes. It is possible that the location of site 113 within a major B-cell antigenic determinant region plays a role in its dynamics, although this is thought to be unlikely.

Site 130

Site 130 is found within the anti-HBc / HBe1 antigenic determinant, a major antibody binding site and a known B-cell epitope. While B-cells are not thought to elicit selection pressue on the core gene, this has not been shown. A recent paper by Ralf Schilling[48] demonstrated the presence of antibodies inside infected hepatocytes, indicating a possible role in eliciting selection pressure.

Site 130 is also highlighted by mulitple analyses, most especially PAML M2a (p88), and the hot spots highlighted by Alexopoulou *et al*[4] (Figure 6:1, p123).
There is also significant support from the histogram data from MacClade, with peaks at site 130 found in both histograms. Site 130 demonstrates the highest degree of changes in the upper histogram (12), and demonstrates high levels of changes in the lower histogram (2 in cases, 2 in controls). Site 130 demonstrates very high significance ($p<0.02$) in the analysis of selection on external branches (p80). Examination of the phylogenetic location of the non-synonymous changes reveals that residue 130 demonstrates between 4 - 6 non-synonymous changes

on external branches (Figure 3:4, p53), and a non-synonymous change on an internal branch (subject 368). Of interest is the observation that site 130 seems to be involved in delineating between case subject 016 and control subject 569 who were HLA class matched (see Figure 3:4, p53)

Site 180

Site 180 is not found within any known immune epitopes, and demonstrates no changes or selection on internal branches, however clearly is important in external analyses of selection ($p = 0.05$). This infers a possible role in transmission selection.

6.1.3 Hypothesis 3: There is a difference in the sites targeted by the immune system

The overall aim of our study was to elucidate sites that may be important to include in a therapeutic vaccine for e-CHB. Therefore we were especially interested in elucidating any sites that were only found to be under positive selection in cases, or only in controls. Our data have excluded this hypothesis.

6.1.4 Transmission problems:

One of the limitations of our study was the inability to determine if a positively selected site was a result of immune pressure or of transmission, as discussed in section 4.4.2 (p94). This represents a controversial issue in virology at present, as the exact mechanisms are unknown. Therefore, further study is required into this to elucidate the precise effects of transmission, and also of immune selection.

A possible solution to this limitation is to conduct a focussed immunological study on the six amino acids that have been highlighted. This will allow a delineation to be made between sites that are selected during transmission and sites that are selected by the host immune system. This could then identify the immune mechanisms responsible for selection at each of the six sites, and allow these mechanisms to be studied in the laboratory.

6.1.5 Other considerations

By structuring our study to match patients based on HLA molecules, we elucidated some interesting results. However due to the wide repertoire of HLA haplotypes contained in our dataset, it is not possible to conclusively determine which residues are associated with which HLA molecule. An interesting further study would be to study all HLA class I molecules in Tongans, thus allowing a greater sample size and greater insight into this dynamic.

Another interesting future study would be to examine if the humoral and cellular immune systems demonstrate differential activity across the HBV genome. This was demonstrated in our data set, with B-cell epitopes exhibiting greater evidence of immune activity than T-cell epitopes. The involvement of the innate immune system also requires further investigation. Our data showed a high incidence of immune activity at known B-cell epitopes, indicating that perhaps the dominant immune response to the core gene (and therefore the nucleocapsid antigen) is the activity of IgG neutralisation of free viral particles. Furthermore, the phenomenon of immunodominance requires investigation, as it is feasible that this is involved here.

The effect of the e antigen is also an interesting hypothesis that requires further investigation and study. Study needs to be undertaken to establish more conclusively if the e antigen is indeed an immunosuppressant, and to elucidate the mechanism of action. A recent study by Alexopoulou et al[49] has suggested that the eAg+ may have an immunosuppressive effect on the production of IgM (anti-HBc), since the eAg does not contain the conformationally dependent dominant B-cell epitope that is found within the core gene.
Following this, both immunological and mathematical models should be constructed to ascertain if there is a positive correlation between the frequency of eAg+ and eAg- clones within a population and the level of selection occurring.
If a positive correlation were found to exist, then further studies would be required to establish if active disease patients demonstrated a higher incidence of frequency dependent selection than inactive disease patients. Such an observation would explain the mechanism for continuing active disease, as well as offering some insight into the mechanism behind the immune deficit observed in our active disease patients.

6.2 Correlation of changes, diversity, epitopes and other studies

Figure 6:1 is included as a summary figure for our study. A legend is included below.

One of the most striking patterns is that of the relationship between known immune epitopes, regions of diversity, and the locations of changes seen in our data set. This is also further correlated with the hotspots identified by Alexopoulou *et al* (indicated by the row "*Diversity = Hotspot*"). This represents an encouraging finding, as it shows that our data are consistent with other analyses.

Figure 6:1 legend:

This figure shows the relationship between sites exhibiting variation, calculated changes, and immune epitopes in three independent studies[4, 50, 51].

(1) The green shading in the rows labelled "Case" and "Controls" reflects any amino acid position that contained more than one amino acid over the entire alignment. This highlights either genetic drift, or selection pressure.

(2) The pink regions in the row labelled " FET Bonferonni adj.p.values" indicate sites that resulted in a p value < 0.05 from a 20 x 2 Fisher's Exact Test, following a Bonferroni correction. This particular test is not highly robust, and so these values were taken as 'indicators of regions of interest', but not necessarily as regions of significance.

(3) The row entitled NS Changes specifies the locations of the non-synonymous changes found within patients.

(4) The coloured regions in the rows labelled 'Ferrari *et al*' reflect the synthetic epitopes that Ferrari *et al* constructed in 1991[50]. The labels refer to the percentage of T-cells that responded to this epitope.

(5) The three rows labelled "Sobao 2402" display results from a study conducted by Sobao *et al* focusing on HLA-A*2402[51].

(6) The row labelled 'Known Immune Epitopes' highlights known epitope regions that have been established in the literature. Red represents T-cell epitopes, whilst

light blue represents B-cell epitopes. Grey regions indicate “major antigenic determinants”.

(7) The dark red discrete regions found in the row entitled “Hotspots - Alexopoulou” represent hot spots of mutation and selection pressure, documented by Alexopoulou *et al* in 2009[4] (24 sites in total). Examination of Figure 6:1 shows that only six of these sites (32, 39, 60, 92, 114, and 174) do not demonstrate variation in our dataset. All other 18 sites correlate to regions of variation and / or a PAML identified site.

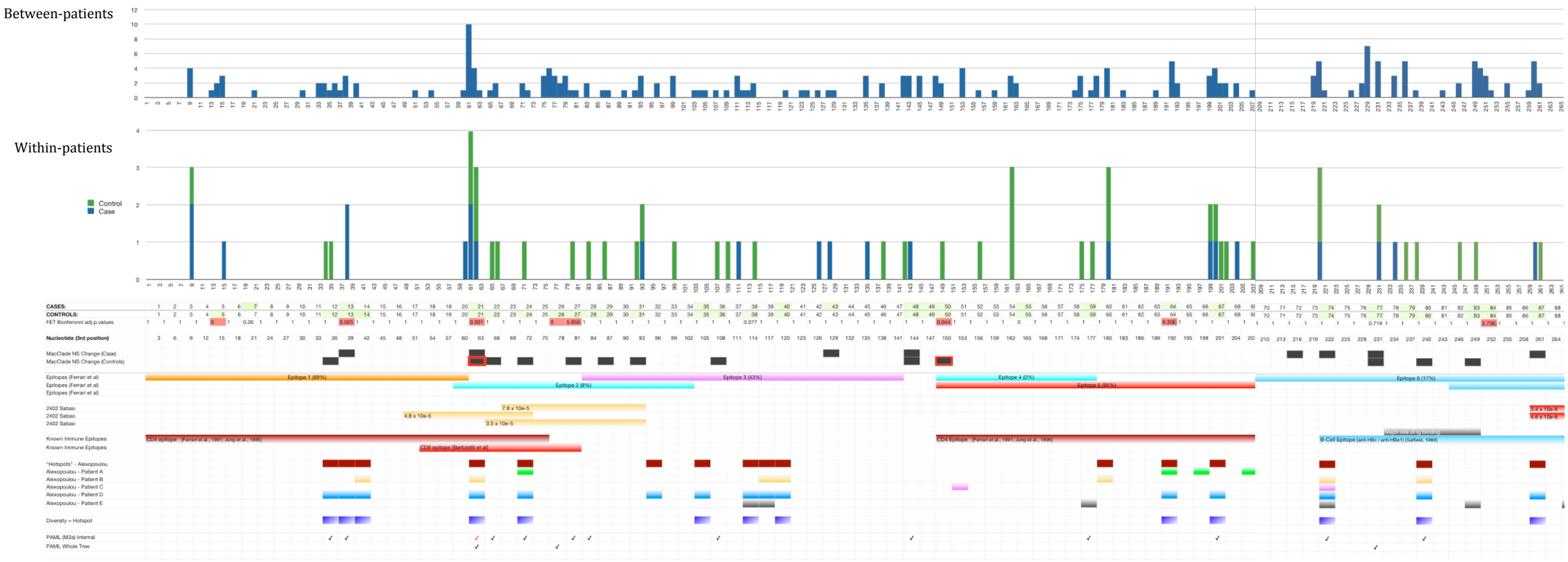
(8) The rows entitled ‘Patient A – Alexopoulou’, ‘Patient B – Alexopoulou’ etc demonstrate five patients from the 2009 study by Alexopoulou *et al*[4]. The coloured regions represent sites that demonstrate mutations / changes found within these patients.

(9) The row entitled ‘PAML M2a’ represents sites that were highlighted by PAML model M2a as being under positive selection. The row entitled PAML Whole Tree represents the six amino acids that were highlighted in the whole tree as being under strong positive selection.

(10), The upper histogram is a representation of the nucleotide changes that have occurred on the whole tree, as calculated by MacClade. This enables analysis of the ‘true number’ of amino acid changes, since a change early in a viral lineage will be amplified by viral replication, causing a false skewing of the numbers. Examination of Figure 6:1 also shows a high correlation between the histogram peaks and the sites identified by PAML M2a. In particular, sites 21, 77, 91, 113, and 130 are highlighted by both analyses.

(11) The lower histogram is a representation of nucleotide changes within patients, and split into cases and controls. There are 44 changes in cases (blue) and 85 changes in controls (green).

Figure 6:1 – Summary Figure




Epitope 7 (17%)
Epitope 8 (26%)
Epitope 9 (47%)
Epitope 10 (26%)
Epitope 11 (0%)
Epitope 12 (0%)
Epitope 13 (4%)
Epitope 14 (4%)
1.7 x 10e-7
< 1 x 10e-8
3.7 x 10e-6
1.2 x 10e-5
1.7 x 10e-5
B-Cell Epitope (anti-HBcHBe2) [Colucci et al., 1988]
Anti-HBcHBe3 determinant [Salfeld et al., 1989]

6.3 Conclusion:

Chronic Active Hepatitis B is a debilitating disease, a major health burden, and as yet no effective cure exists. Viral persistence is a poorly understood phenomenon, and HBV is no exception. Clearly there is much to learn about the operation of the human immune system and its interactions with viral pathogens.
This study aimed to further elucidate the host – pathogen dynamics, with specific focus on mutational repertoires found within the virus.

Hepatitis B viral infection presents an interesting paradox, whereby the disease is both controlled and exacerbated by a single agent, the immune system. Examination of diversity, selection pressure and mutational repertoires within viral clones extracted from chronically infected patients demonstrated no major deficit within active disease patients, however a subtle deficit is not excluded. We have also proposed that frequency dependent selection is a likely mechanism for this deficit; the immuno-suppressive effects of the eAg thus require further study, perhaps at both an immunological and a mathematical level.

Our study has shown that controlling for HLA haplotype status is indeed beneficial, and provides valuable insight into optimum requirements for sample sizes and study design. Further studies could incorporate a similar number of clones per patient, but a greater number of patients.

In conclusion, there is still much to learn about chronic hepatitis B infection. This study is one small step in the overall goal of creating a therapeutic vaccine to aid those already affected.

7 References :

1. Abbott, W.G.H., *Personal Communication - Study Parameters*, B. Warner, Editor. 2008: Auckland.
2. Seeger, C. and W.S. Mason, *Hepatitis B virus biology.* Microbiology and Molecular Biology Reviews, 2000. **64**(1): p. 51-68.
3. Chang, M.-H., *Hepatitis B virus infection.* Seminars in Fetal and Neonatal Medicine, 2007. **12**(3): p. 160-167.
4. Alexopoulou, A., *Mutants in the precore, core promoter, and core regions of Hepatitis B virus, and their clinical relevance.* Annals of Gastroenterology, 2009. **22**(1): p. 13-23.
5. Bozkaya, H., B. Ayola, and A.S.F. Lok, *High rate of mutations in the hepatitis B core gene during the immune clearance phase of chronic hepatitis B virus infection.* Hepatology, 1996. **24**(1): p. 32-37.
6. Bertoletti, A. and A.J. Gehring, *The immune response during hepatitis B virus infection.* Journal of General Virology, 2006. **87**(6): p. 1439-1449.
7. Hoofnagle, J.H., G.M. Dusheiko, and L.B. Seeff, *Seroconversion from hepatitis B e antigen to antibody in chronic type B hepatitis.* Annals of Internal Medicine, 1981. **94**(6): p. 744-748.
8. Akarca, U.S. and A.S.F. Lok, *Naturally occurring hepatitis B virus core gene mutations.* Hepatology, 1995. **22**(1): p. 50-60.
9. Hadziyannis, S.J. and D. Vassilopoulos, *Hepatitis B e antigen-negative chronic hepatitis B.* Hepatology, 2001. **34**(4 I): p. 617-624.
10. Lavanchy, D., *Hepatitis B virus epidemiology, disease burden, treatment, arid current and emerging prevention and control measures.* Journal of Viral Hepatitis, 2004. **11**(2): p. 97-107.
11. Lok, A.S.F. and B.J. McMahon, *Chronic hepatitis B.* Hepatology, 2001. **34**(6): p. 1225-1241.
12. Beasley, R.P., et al., *Hepatocellular carcinoma and hepatitis B virus.* Lancet, 1981. **2**(8256): p. 1129-1132.
13. Health, M.O. (2004) *Key results of the 2002/03 New Zealand Health Survey.* A Portrait of Health.
14. Michel, M.L. and M. Mancini-Bourgine, *Therapeutic vaccination against chronic hepatitis B virus infection.* Journal of Clinical Virology, 2005. **34**(SUPPL. 1).
15. Cote, P.J., et al., *Temporal pathogenesis of experimental neonatal woodchuck hepatitis virus infection: Increased initial viral load and decreased severity of acute hepatitis during the development of chronic viral infection.* Hepatology, 2000. **32**(4 I): p. 807-817.
16. Hainaut, P. and P. Boyle, *Curbing the liver cancer epidemic in Africa.* The Lancet. **371**(9610): p. 367-368.
17. Wieland, S.F., et al., *Interferon prevents formation of replication-competent hepatitis B virus RNA-containing nucleocapsids.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(28): p. 9913-9917.
18. Wieland, S.F. and F.V. Chisari, *Stealth and cunning: Hepatitis B and hepatitis C viruses.* Journal of Virology, 2005. **79**(15): p. 9369-9380.

19. Kramvis, A., M. Kew, and G. FranÁois, *Hepatitis B virus genotypes.* Vaccine, 2005. **23**(19): p. 2409-2423.
20. Chotiyaputta, W. and A.S.F. Lok, *Hepatitis B virus variants.* Nature Reviews Gastroenterology and Hepatology, 2009. **6**(8): p. 453-462.
21. Berquist, K.R., J.M. Peterson, and B.L. Murphy, *Hepatitis B antigens in serum and liver of chimpanzees acutely infected with hepatitis B virus.* Infection and Immunity, 1975. **12**(3): p. 602-605.
22. Huo, T.I., et al., *Hepatitis B virus X mutants derived from human hepatocellular carcinoma retain the ability to abrogate p53-induced apoptosis.* Oncogene, 2001. **20**(28): p. 3620-3628.
23. Milich, D.R., et al., *Is a function of the secreted hepatitis B e antigen to induce immunologic tolerance in utero?* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(17): p. 6599-6603.
24. Maini, M.K., et al., *The role of virus-specific CD8+ cells in liver damage and viral control during persistent hepatitis B virus infection.* Journal of Experimental Medicine, 2000. **191**(8): p. 1269-1280.
25. Thimme, R., et al., *CD8+ T cells mediate viral clearance and disease pathogenesis during acute hepatitis B virus infection.* Journal of Virology, 2003. **77**(1): p. 68-76.
26. Penna, A., et al., *Cytotoxic T lymphocytes recognize an HLA-A2-restricted epitope within the hepatitis B virus nucleocapsid antigen.* Journal of Experimental Medicine, 1991. **174**(6): p. 1565-1570.
27. Reignat, S., et al., *Escaping high viral load exhaustion: CD8 cells with altered tetramer binding in chronic hepatitis B virus infection.* Journal of Experimental Medicine, 2002. **195**(9): p. 1089-1101.
28. Nowak, M.A., et al., *Viral dynamics in hepatitis B virus infection.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(9): p. 4398-4402.
29. Bertoletti, A., et al., *Cytotoxic T lymphocyte response to a wild type hepatitis B virus epitope in patients chronically infected by variant viruses carrying substitutions within the epitope.* Journal of Experimental Medicine, 1994. **180**(3): p. 933-943.
30. Rodrigo, A.G., *Personal Communication - Viral Evolution*, B. Warner, Editor. 2009: Auckland.
31. Long, C., H. Qi, and S.H. Huang, *Mathematical modeling of cytotoxic lymphocyte-mediated immune response to hepatitis B virus infection.* Journal of Biomedicine and Biotechnology, 2008. **2008**(1).
32. Lau, G.K.K., et al., *Resolution of chronic hepatitis B and anti-HBs seroconversion in humans by adoptive transfer of immunity to hepatitis B core antigen.* Gastroenterology, 2002. **122**(3): p. 614-624.
33. Wainwright, R.B., B.J. McMahon, and T.R. Bender, *Prevalence of hepatitis B virus infection in Tonga: Identifying high risk groups for immunization with hepatitis B vaccine.* International Journal of Epidemiology, 1986. **15**(4): p. 567-571.
34. Wilson, N., et al., *The effectiveness of the infant hepatitis B immunisation program in Fiji, Kiribati, Tonga and Vanuatu.* Vaccine, 2000. **18**(26): p. 3059-3066.
35. Abbott, W.G.H., et al., *Low-cost, simultaneous, single-sequence genotyping of the HLA-A, HLA-B and HLA-C loci.* Tissue Antigens, 2006. **68**(1): p. 28-37.

36. Ross, H.A. and A.G. Rodrigo, *Immune-Mediated Positive Selection Drives Human Immunodeficiency Virus Type 1 Molecular Variation and Predicts Disease Duration.* J. Virol., 2002. **76**(22): p. 11715-11720.
37. Agencourt-Bioscience, *Agencourt CleanSEQ Dye-terminator removal protocol*. 2006, Beckman Coulter: Beverly, Massachusetts. p. 12.
38. Drummond AJ, A.B., Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A, *Geneious*. 2009, Biomatters Ltd.
39. Swofford, D., *PAUP**, Sinauer Associates: Sunderland, Massachusetts. p. Phylogenetic Analysis Program.
40. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-2948.
41. Guindon S, G., *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Systematic Biology, 2003. **52**(5): p. 696-704.
42. Lim, S.G., et al., *Viral Quasi-Species Evolution During Hepatitis Be Antigen Seroconversion.* Gastroenterology, 2007. **133**(3): p. 951-958.
43. Norder, H., et al., *Genetic diversity of hepatitis B virus strains derived worldwide: Genotypes, subgenotypes, and HBsAg subtypes.* Intervirology, 2004. **47**(6): p. 289-309.
44. Marinos, G., et al., *Tumor necrosis factor receptors in patients with chronic hepatitis B virus infection.* Gastroenterology, 1995. **108**(5): p. 1453-1463.
45. Abbott, W.G.H., et al., *Associations between HLA class I alleles and escape mutations in the hepatitis B Virus core gene in New Zealand-resident tongans.* Journal of Virology, 2010. **84**(1): p. 621-629.
46. Morgenstern, J.v.N.O., *Theory of Games and Economic Behaviour*. 1944, Princeton: Princeton University Press.
47. Riedl, P., et al., *Distinct, cross-reactive epitope specificities of CD8 T cell responses are induced by natural hepatitis B surface antigen variants of different hepatitis B virus genotypes.* Journal of Immunology, 2006. **176**(7): p. 4003-4011.
48. Schilling, R., et al., *Endocytosis of hepatitis B immune globulin into hepatocytes inhibits the secretion of hepatitis B virus surface antigen and virions.* Journal of Virology, 2003. **77**(16): p. 8882-8892.
49. Alexopoulou, A., et al., *Core mutations in patients with acute episodes of chronic HBV infection are associated with the emergence of new immune recognition sites and the development of high IgM Anti-HBc index values.* Journal of Medical Virology, 2009. **81**(1): p. 34-41.
50. Ferrari, C.B., A. Penna, A. Cavalli, A. Valli, A. Missale, G. Pilli, M. Fowler, P. Gluberti, T. Chisari, FV. Fiaccadori, F., *Identification of Immunodominant T Cell Epitopes of the HBV Nucelocapsid Antigen.* Journal of Clinical Investigation, 1991. **88**: p. 214-222.
51. Sobao, Y., et al., *Identification of hepatitis B virus-specific CTL epitopes presented by HLA-A*2402, the most common HLA class I allele in East Asia.* Journal of Hepatology, 2001. **34**(6): p. 922-929.

8 Appendices:

8.1 Materials:

8.1.1 Agarose Gel Electrophoresis Buffers

10x Agarose Gel Sample Loading Buffer		250mg bromophenol blue, dissolved in 33ml 150mM Tris (pH 7.6), plus 60ml glycerol and 7ml H_2O
50x TAE		242g Tris dissolved in 1000ml H_2O, 100ml 0.5M Na_2 EDTA (pH 8.0), 57.1 Glacial Acetic Acid
Agarose		

8.1.2 Antibiotics

Ampicillin		Sodium ampicillin stock (Sigma A9518) 100mg/ml, 1.5g of sodium ampicillin to a sterile red-top 15ml tube, add 15 ml sterile ddH_2O. Dissolve at room temperature and aliquot into eppendorfs. Store at -20℃.

8.1.3 Bacterial Cell Culture

SOC Media		2.0g Bacto-tryptone, 0.5g Bacto-yeast extract, 1ml 1M NaCl, 0.25ml 1M KCl, 1ml 2M Mg^{2+} stock (filter sterilised), 1ml 2M glucose (filter sterilised), pH 7.0
2M Mg^{2+} Stock		20.33g $MgCl_2 \bullet 6H_2O$, 24.65g $MgSO_4 \bullet 7H_2O$, ddH_2O to 100ml, Filter Sterilise
LB Agar		17.5g LB Agar (Sigma), 500ml ddH_2O. Autoclave
Ampicillin Inoculum		30µL Ampicillin, 170µL ddH_2O.
Escherichia Coli		Dh5α.
Terrific Broth		2.2212g Glycerol, 10.472g Terrific Broth Medium, 220mL ddH2O

8.1.4 Cloning Buffers and Solutions

40% PEG Working Solution		40g PEG 8000, 1mL 1M $MgCl_2$, 100mL dH_2O, Filter
TE Buffer		1ml 1M Tris HCl pH 8.0, 0.2ml 0.5M EDTA, make up to 100ml with ddH_2O, autoclave
5M NaCl		29.22g NaCl dissolved in 100ml ddH_2O, autoclave
PEG Cleaning Solution		36µL TE Buffer, 36µL 5M NaCl, 63µL 40% PEG Working Solution
A-Tailing Solution		2.34µL ddH_2O, 0.88µL 10x Buffer, 0.8µL dATP, 0.33µL $MgCl_2$, 0.45µL Taq Polymerase
X-GAL		Promega, Madison, Wisconsin, USA, CAT# V394A
IPTG		Promega, Madison, Wisconsin, USA

8.1.5 DNA Extraction Buffers and Solutions

Phenol pH 7.6		0.5g 8-Hydroxyquinolone, 500ml 1M TrisHCl pH 8.0
Phenol Chloroform		Sigma, Scharlau
Resuspension Buffer		24ml 0.5M EDTA pH 8.0, 37.5ml 1M NaCl, make up to 500ml with ddH_2O
SDS		Ackross Chemicals
Sucrose Lysis Buffer		109.5g Sucrose, 5ml 1M $MgCl_2$, 10g Triton-X, 10ml 1M TrisHCl pH 7.5, make up to 1L with ddH_2O, Autoclave

8.1.6 Enzymes

T4 DNA Ligase		Promega, Madison, Wisconsin, USA

Taq Polymerase		Prof. John Fraser, Auckland, NZ
Accuprime Taq		Promega, Madison, Wisconsin, USA

8.1.7 General Buffers

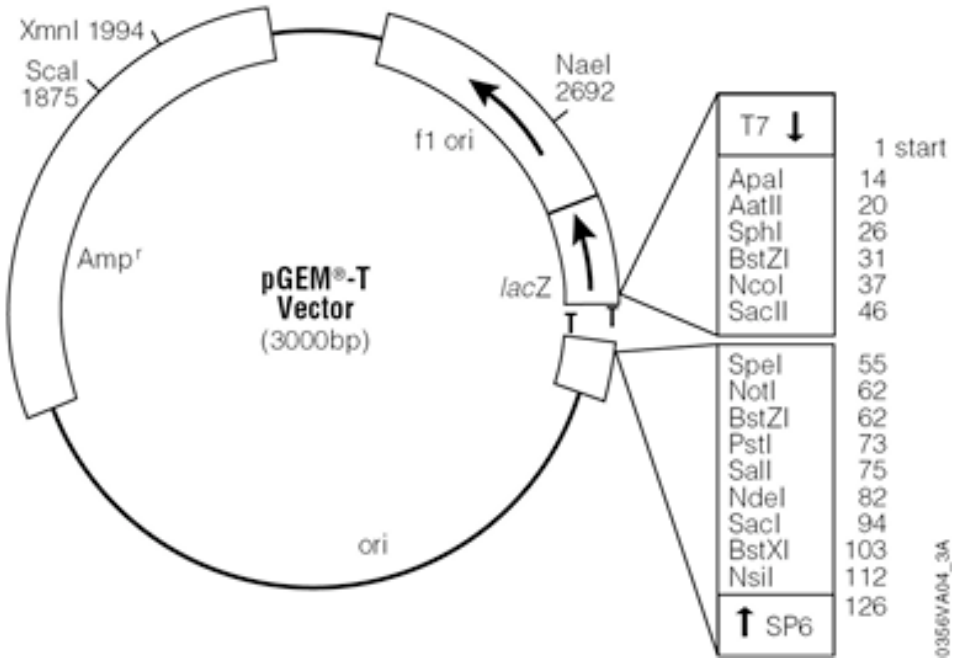
10x Ammonium Sulfate Buffer		160mM $(NH_4)_2SO_4$, 670mM TrisHCl pH 8.4, 0.1% Tween20
1M TrisHCl	pH 7.4	121.14g Tris, 800ml H2O, 70ml HCl
	pH 7.6	121.14g Tris, 800ml H2O, 60ml HCl
	pH 8.0	121.14g Tris, 800ml H2O, 42ml HCl

8.1.8 Molecular Weight Markers

250 bp ladder,		Invitrogen [Cat #: 10596-013]

8.1.9 Plasmids

pGEM-T A3600		Promega, Madison, Wisconsin, USA



8.1.10 Polymerase Chain Reaction Components

dNTP / dATP / dTTP		100mM dNTPs (Promega, Madison, WI)
5x Sucrose		4.25g Sucrose, 10mL H_20 for injections, Stored at -20℃
Primer Solutions		Invitrogen, Stored as 100pM in TE Buffer
25mM $MgCl_2$		5.08g $MgCL_2•6H_2O$, 800ml ddH_2O, adjust to 1L, autoclave

8.1.11 Sequencing Solutions

5x Sequencing Buffer (CleanSeq)		Agencourt Biosciences Corp., Beverley, MA
ABI Prism BigDye Terminator (BDT) v3.1 cycle sequencing mix		Applied Biosystems, Foster City, CA

8.1.12 Viral Extraction

Roche Viral Extraction Kit		Roche Applied Science, Mannheim, Germany

8.2 Clones that were excluded:

For a variety of reasons, it was necessary to exclude some clones from the final analysis. The reasons for exclusion are summarised in the table below.

Reason	Clone
Poor sequence quality	290I, 308A
Multiple sequences contained in file	008C, 249B, 277E, 368M
Insertion / Deletion in Core ORF	249P, 318F, 337H, 539A
Premature Stop Codon	247F, 277B, 277C, 305A, 305B, 305F, 305G, 305H, 305I, 305K, 305O, 305P, 305Q, 337A, 553A,
Possible contaminant (did not cluster with other clones from patient)	008J, 249G, 290H, 277H, 290K, 290M, 368K,
Failed to Transform	008E, 029C, 249F, 250G, 305E
Mislabeled	233F

8.3 Model M1a Output data

Subject		Negative	Neutral
008	*p*	0.7412	0.2588
	w	0	1
016	*p*	0.99999	0.00001
	w	0	1
029	*p*	0.87863	0.12137
	w	0	1
249	*p*	0.66078	0.33922
	w	0	1
250	*p*	0.65087	0.34913
	w	0	1
290	*p*	0.99999	0.00001
	w	0.06478	1
305	*p*	0.53345	0.46655
	w	1	1
308	*p*	0.99999	0.00001
	w	0	1
318	*p*	0.99999	0.00001
	w	0.90135	1
337	*p*	0.19366	0.80634
	w	0	1
365	*p*	0.99999	0.00001
	w	0.10883	1
368	*p*	0.28373	0.71627
	w	0	1
413	*p*	0.28195	0.71805
	w	0	1
459	*p*	0.99999	0.00001
	w	0.4552	1
553	*p*	0.49836	0.50164
	w	0	1

8.4 Likelihood ratio test for internal selection pressure

Patient	LnL 1a	LnL 2a	ABS(2*Ln1a-Ln2a)	p-value	Bonferroni Corrected
008	-846.459678	-842.714157	7.4910	0.0236	0.3540
016	-751.940168	-751.940466	0.0006	0.9997	14.9955
029	-772.313228	-764.001702	16.6231	0.0002	0.0030
249	-915.692277	-912.478729	6.4271	0.0402	0.6030
250	-811.535926	-809.798067	3.4757	0.1758	2.6370
290	-790.257302	-790.257314	0.0000	1.0000	15.0000
305	-858.268222	-857.594568	1.3473	0.5098	7.6470
308	-767.546583	-767.546316	0.0005	0.9997	14.9955
318	-764.990356	-764.994256	0.0078	0.9961	14.9415
337	-791.974596	-789.911158	4.1269	0.1270	1.9050
365	-781.863689	-781.863516	0.0003	0.9998	14.9970
368	-864.499795	-864.070618	0.8584	0.6510	9.7650
413	-829.712067	-827.494701	4.4347	0.1088	1.6320
553	-768.447685	-768.447417	0.0005	0.9997	14.9955
569	-760.836087	-760.83672	0.0013	0.9993	14.9895

Table 8:1- Likelihood Ratio Test results, showing log likelihood values, the difference, the pre-corrected p-value and the post-corrected p-value.

8.5 Likelihood ratio test for external selection pressure

LnL 1a	LnL 2a	|2*(LnL1a-LnL2a)|	p
-2027.385887	-1990.630925	73.509924	0.000002

8.6 Likelihood ratio test for entire tree

LnL 1a	LnL 2a	|2*(LnL1a-LnL2a)|	p
3280.623951	3261.451848	38.344206	0.0000003

8.7 Character changes

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
16	551	1		184		
365	548	4		184		
553	548	4		182	2	
569	550	2		183	1	

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
29	549	2	1	183	1	
305	532	20		169	15	

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
249	533	14	5	174	6	4
290	545	7		183	1	
368	538	13	1	175	8	1

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
250	543	9		180	4	
8	540	9	3	178	4	2

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
308	549	3		184		
413	542	9	1	177	6	1

Patient	Nucleotide			Amino Acid		
	0 changes	1 changes	2 changes	0 changes	1 changes	2 changes
318	549	3		182	2	
337	545	7		180	4	

8.8 Points of interest from the Phylograms

Detailed examination of the phylograms reveals multiple points of interest.

Patient 318 and 290:

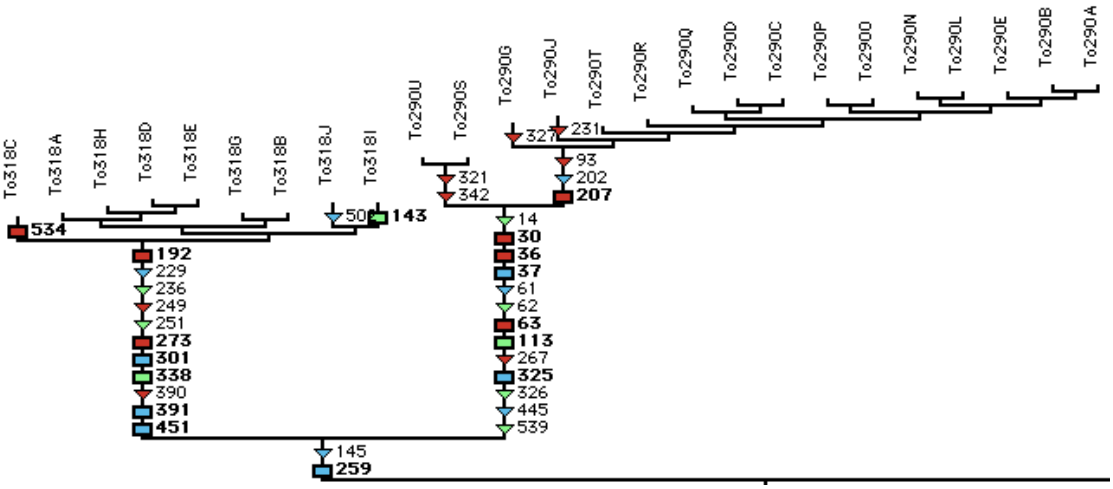


Figure 8:1 – Reconstruction by MacClade of a section of the core gene phylogram. Showing patients 318 and 290, and nucleotide changes coloured by position in the codon. (1st position = blue; 2nd position = green; 3rd position = red).

Close inspection of the core gene phylogram revealed this region highlighted above. Of note here is the recent divergence of two patients, one who is a 'case' (318) and the other who is a 'control' (290). The above diagram elucidates which sites on each branch have changed, and colours them according to their position in the codon. Interestingly, the sites highlighted correspond to sites that are highlighted repeatedly in other analyses.

Cases – residues 64, 77, 79, 83, 84, 91, 101, 113, 130, 131, 151; Controls – residues 5, 10, 12, 13, 21, 38, 89, 109, 149, 180.

Patient 569 versus patient 016:

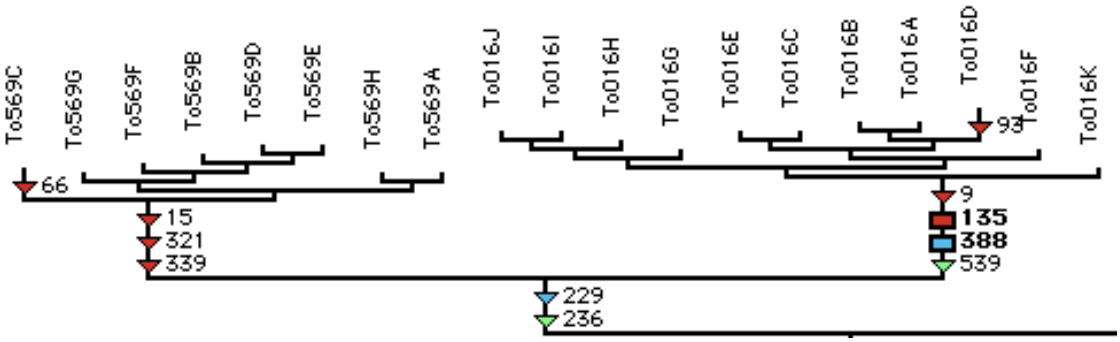


Figure 8:2 – Reconstruction by MacClade of a section of the core gene phylogram. Showing patients 016 and 569, and nucleotide changes coloured by position in the codon. (1st position = blue; 2nd position = green; 3rd position = red).

Patient 569 (inactive) and patient 016 (active) were HLA class matched, as described earlier. Therefore, their clustering together on the phylogram is of interest. It can be seen from the diagram above that the differences between case and control are small and mostly occur at third position sites. The notable exception to this is sites 388 and 539, which differentiate 016 from 569. These sites are of interest, as they correspond to residues 130 and 180, which have been identified in other analyses as being of significance.

8.9 Patient by Patient tree diversity comments

Cases – Active disease

Patient 318:

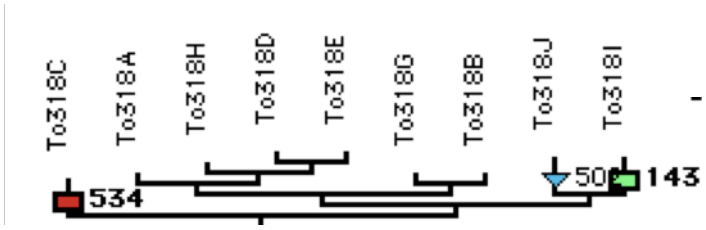


Figure 8:3
Low levels of diversity are seen within this patient. Only three changes are seen, with one for each position in the codon. The nucleotide changes correspond to residues 48, 167, and 178.

Patient 250:

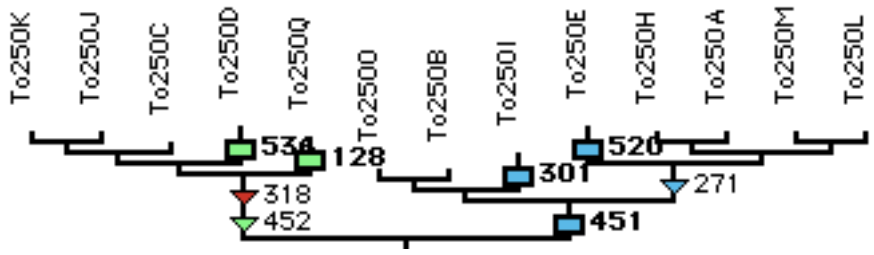


Figure 8:4

There is a high level of diversity seen within patient 250. Eight changes are seen, with seven being at either the first or second position. Five of these changes are unique to this clade. The changes correspond to residues 43, 91, 101, 106, 151, 174 and 179.

Patient 016:

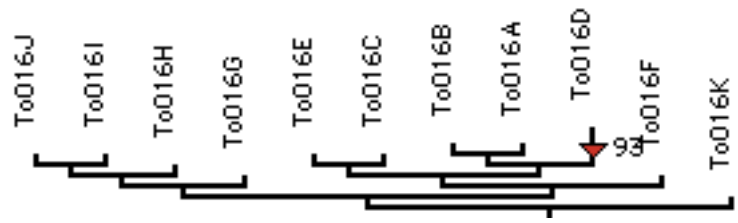


Figure 8:5

Very low diversity is seen within this patient, with only one change (site 93, residue 31), likely to be due to 'wobble'.

Patient 249:

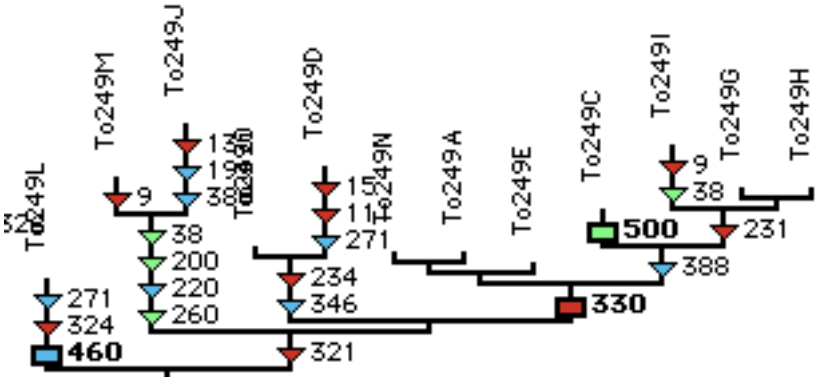


Figure 8:6

Patient 249 demonstrates the highest level of diversity amongst all patients. The above figure shows 23 changes, 10 of which are third position changes. The above changes correspond to residues 3, 5, 13, 37, 45, 67, 74, 77, 78, 87, 91, 107, 108, 110, 116, 129, 130, 154, and 167.

Patient 308:

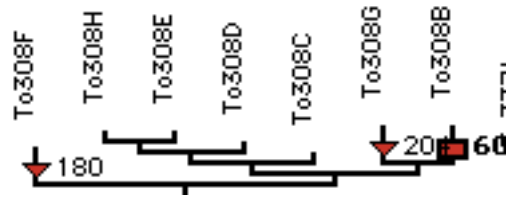


Figure 8:7

Patient 308 show very little variation, with three changes, all found at the third position, corresponding to residues 20, 60, and 68.

Patient 029:

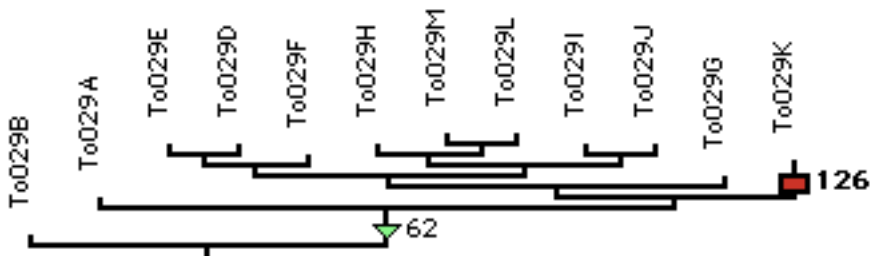


Figure 8:8

Very little diversity is seen within patient 029 also, with only two changes. Site 62 corresponds to residue 21. Site 126 corresponds to residue 42.

Controls – Inactive disease

Patient 413:

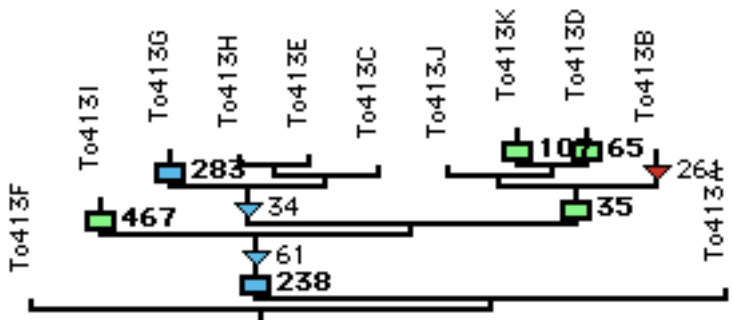


Figure 8:9

There is a high level of diversity seen within patient 413, with nine changes. Only one of these changes is likely to be 'wobble'. An analysis of the amino acid changes for this clade reveals that there are 8 amino acid changes.
These changes correspond tor residues 12, 21, 22, 36, 80, 87, 95, and 156.

Patient 553:

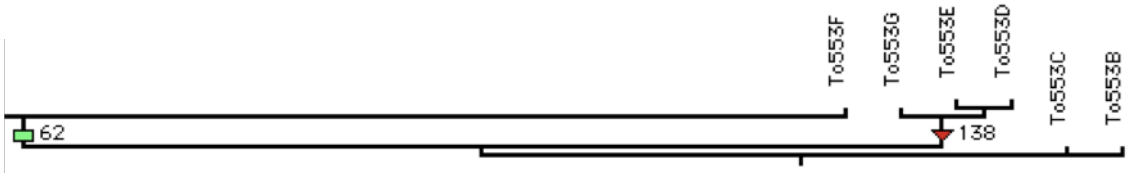


Figure 8:10

Is found at the root of the ML tree, and demonstrates a low level of variation (two changes), corresponding to residues 21 and 46.

Patient 305:

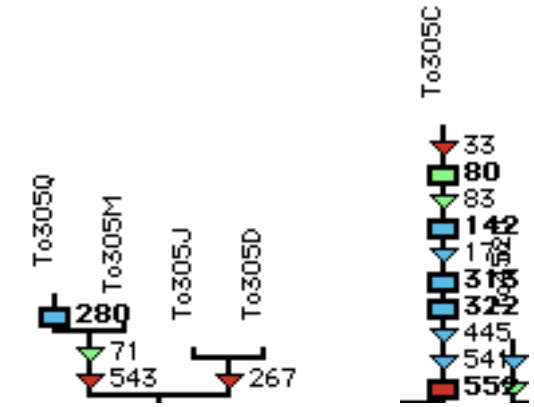


Figure 8:11

Patient 305 demonstrates moderate diversity within four clones, with the fifth clone (305C) not clustering within the clade. Clone 305C demonstrates a high level of changes and seems atypical for its kind. Patient 305 also often contained premature stop codons, requiring these clones to be excluded.
The above changes (not including 305C) correspond to residues 24, 89, 94, 181. The changes found within 305C correspond to residues 11, 27, 28, 48, 59, 105, 108, 149, 181, and 184.

Patient 290:

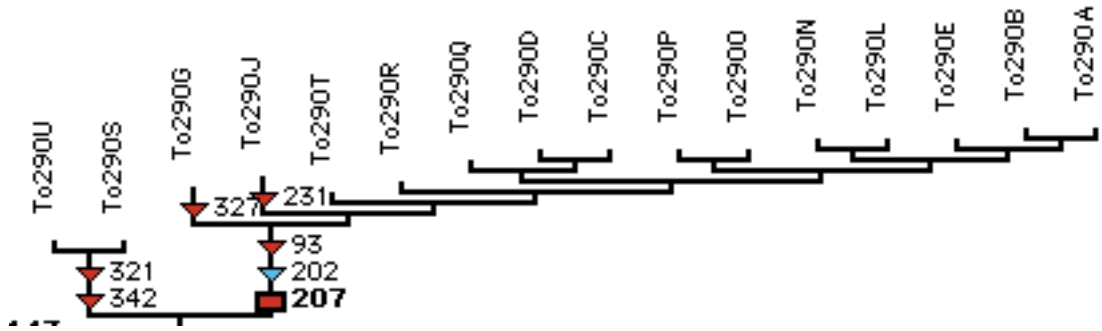


Figure 8:12

There is some variation found within patient 290, however this is mostly found between two clades and at the third position. There are seven sites that change in total. These changes correspond to residues 31, 68, 69, 77, 107, 109, and 114.

Patient 459:

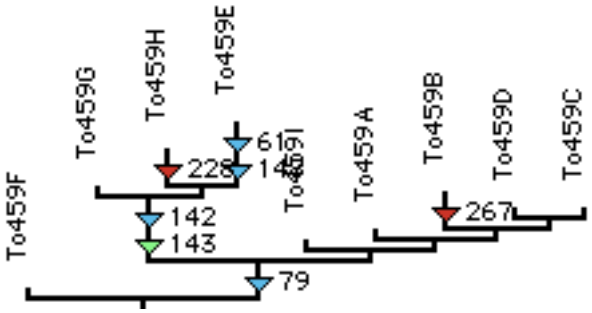


Figure 8:13

Demonstrates a moderate level of variation, with seven sites varying. Only two of these sites are likely to be wobble. These changes correspond to residues 21, 27, 48, 49, 76, and 89.

Patient 368:

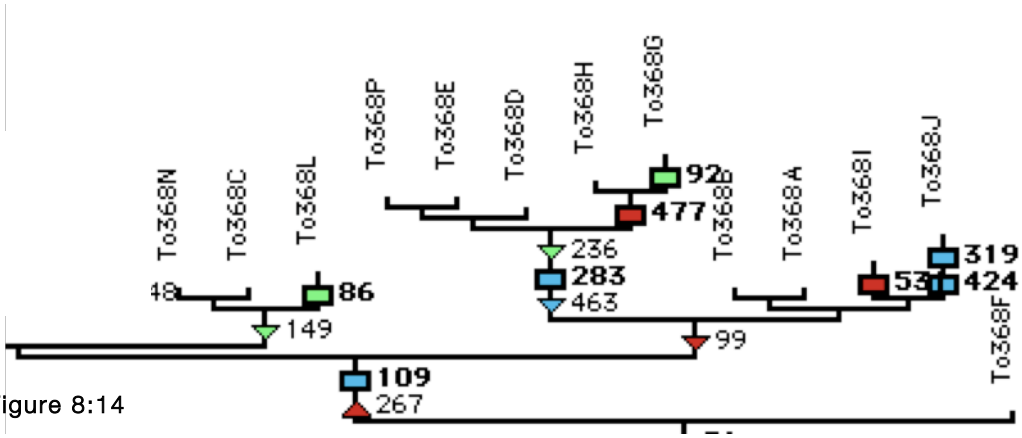


Figure 8:14

Exhibits a high amount of variation, as evidenced by the figure above. There are multiple clades within this patient, as well as changes occurring between most clones. There are 13 changes in total, with four likely to be wobble. The changes correspond to residues 29, 31, 33, 37, 50, 79, 89, 95, 107, 142, 155, 159, and 177. Patient 368 is a recent seroconvertant.

Patient 569:

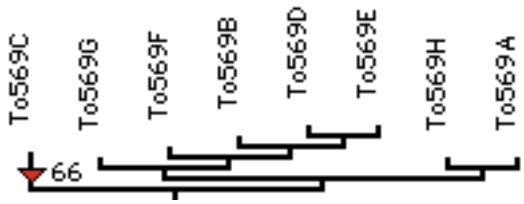


Figure 8:15

Demonstrates a low level of variation, with only one site undergoing any change (residue 22)

Patient 365:

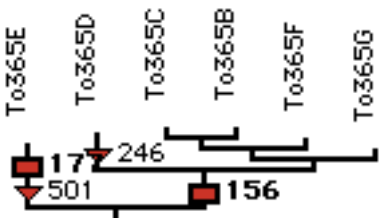


Figure 8:16

Exhibits four changes, all of which occur at the third position and are likely to be ‘wobble’. These correspond to residues 6, 52, 82, and 167.

Patient 008:

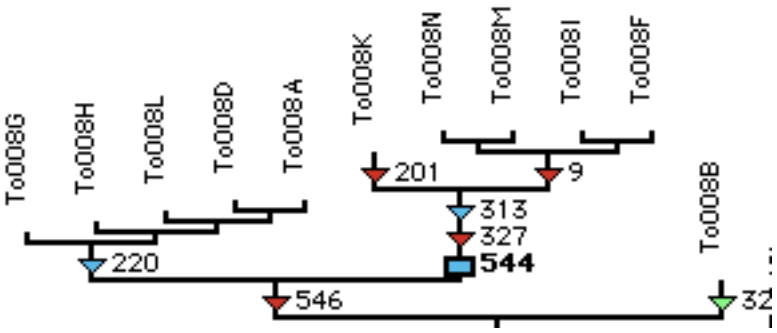


Figure 8:17

Demonstrates three main clades within the patient. There are eight changes in total, with half being at the third position. The changes correspond to residues 3, 11, 67, 74, 105, 109, and 182.

Patient 337:

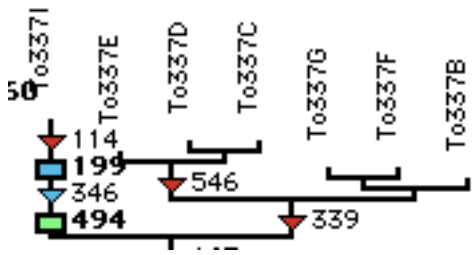


Figure 8:18

Shows a moderate level of variation, with six changes seen. Half of these changes are at the third position. The changes correspond to residues 38, 67, 113, 116, 165, and 182.

8.10 Residues under variation within clades

The table below highlights which sites appear in multiple patients, using the following colour coding: *Green* – 1 patient shows activity at the site in both case and control; *Yellow* – 2 patients show activity at this site only in controls; *Purple* – 2 patients show activity at this site only in cases; *Blue* – 3 or more patients show activity at this site.

Table 8:2 - Location of all nucleotide changes, shown at the amino acid level. *X* – change at position 1 or 2. *O* – change at position 3.

	Cases							*Controls*								
	016	*318*	*250*	*249*	*308*	*029*		*413*	*553*	*305*	*290*	*459*	*368*	*569*	*365*	*008*	*337*
3				o												o	
5				o													
6															o		
11										o						x	
12								x									
13				x													
20					o												
21						x		x	x			x					
22								x						o			
24										x							
27										x		x					
28										x							
29													x				
31	o										o		x				
33													o				
36								x									
37				o									x				
38																	o
42						o											
43			x														
45				o													
46									o								
48		x								x		x					
49												x					
50													x				
52															o		
59										x							
60					o												
67				x												o	x
68					o						x						
69											o						
74				x												x	
76												o					
77				o							o						
78				o													
79													x				
80								x									
82															o		
87				x				o									
89										o		o	o				
91			x	x													
94										x							

95				x			x			
101		x								
105					x				x	
106		o								
107			o			o	x			
108			o		x					
109						o			x	
110			o							
113										o
114						o				
116			x							x
129			x							
130			x							
142							x			
149					x					
151		x								
154			x							
155							x			
156				x						
159							o			
165										x
167	x		x					o		
174		x								
177							o			
178	o									
179		x								
181					o					
182									o	o
184					o					

8.11 Clade Comparisons

MacClade inter-patient analysis

Of particular interest was whether the changes seen in each clade were unique to that clade (patient specific) or were unique to cases, or to controls. To that end we examined the changes that occurred in each clade. Here we focus on the two clades from both the controls (290 and 365) and the cases (318 and 308) with the highest levels of changes, as calculated by MacClade.

Patient 290 – Inactive disease (control)

There are multiple observations to be made from this analysis. There are four first position changes (sites 37, 61, 325, 445) indicating a non-synonymous change, and five second position changes (14, 62, 113, 326, 539), which may or may not contribute to a non-synonymous change. There is a noticeably higher level of transversions than transitions (9 : 4).
Sites **14**, 36, 62, 326, **445**, and 539 (Residues 5, 12, 21, 109, 149, and 180 respectively) also demonstrate convergence with states outside this clade, indicating that these changes are not unique to this clade. Residues in bold indicate a first position change. When the analysis was restricted to only controls, only sites 326, 445 and 539 were found as being found elsewhere, inferring that they may be changes that are specific to controls.
The sites labeled as 'reversals' are of interest. A reversal indicates that this position has changed from state 1 → state 2 earlier in the lineage, and now has changed from state 2 → state 1 again. Site 267 is a third position site, and is likely to be an indication of 'wobble'. Site 61 (residue 21) however is a first position site, thus potentially demonstrating a hyper-variable region. It is worth noting that site 61 is also a transversion. Site 61 is also demonstrates a change in patient 413 and 305.

Patient 365 – Inactive disease (control)

In this clade we see eight nucleotides (40, 61, 199, 220, 229, 313, 439, and 445) at the first position undergoing change, and only one site (539) at the second position showing change. These sites correspond to residues 14, 21, 67, 74, 77, 105, 147, and 149, respectively. All sites are non-unique. When the analysis was restricted to control patients only, sites 40 and 439 were classified as unique. This infers that these sites are not found in any of the other control patients. Sites 61, 435, and 510 demonstrate reversal. 435 and 510 are third position sites and therefore likely to be wobble. Of note is site 61 (residue 21), which is a first position site and a transversion. This again infers hyper-variability at this site.
The transversion : transition ratio for this branch is 6 : 8

Table 8:3 – Patient 290

Site	Position	Within clade	Other clades	Ts/Tv
14	2		State at this node convergent with a state found outside this clade	Transversion
30	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
36	3	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion
37	1	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
61	1		State at this node found within an ancestor, thus representing a reversal	Transversion
62	2		State at this node convergent with a state found outside this clade	Transversion
63	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
113	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
267	3		State at this node found within an ancestor, thus representing a reversal	Transition
325	1	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
326	2		State at this node convergent with a state found outside this clade	Transversion
445	1		State at this node convergent with a state found outside this clade	Transition
539	2		State at this node convergent with a state found outside this clade	Transition

Table 8:4 – Patient 365

Site	Position	Within clade	Other clades	Ts/Tv
40	1		State at this node convergent with a state found outside this clade	Transversion
61	1		State at this node found within an ancestor, thus representing a reversal	Transversion
199	1		State at this node convergent with a state found outside this clade	Transition
220	1		State at this node convergent with a state found outside this clade	Transition
229	1		State at this node convergent with a state found outside this clade	Transversion
313	1		State at this node convergent with a state found outside this clade	Transition
327	3		State at this node convergent with a state found outside this clade	Transversion
339	3		State at this node convergent with a state found outside this clade	Transversion
435	3		State at this node found within an ancestor, thus representing a reversal	Transition
439	1		State at this node convergent with a state found outside this clade	Transversion
445	1		State at this node convergent with a state found outside this clade	Transition
510	3		State at this node found within an ancestor, thus representing a reversal	Transition
539	2		State at this node convergent with a state found outside this clade	Transition
543	3		State at this node convergent with a state found outside this clade OR State at this node found within an ancestor, thus representing a reversal	Transition

Patient 308 – active disease (case):

Here we see a predominance of third position changes (5 / 11), with only 3 / 11 changes occurring at the first position (leading to a NS change). These are sites 226, 250, and 466 (residues 76, 84, and 156, respectively). The three second position changes (251, 404, and 539; residues 84, 135, and 180, respectively) may contribute to NS changes. Three sites are found to be unique to this clade (21, 54 and **226**). Site 226 is a first position change. When the analysis was restricted to cases only, sites 466 and 539 were the only sites that were classified as non-unique. This could infer that these changes are confined to the case population only.
The ratio of transversions to transitions is 5 : 6

Table 8:5 – Patient 308

Site	Position	Within clade	Other clades	Ts/Tv
21	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
54	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
111	3	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion
201	3		State at this node convergent with a state found outside this clade	Transition
226	1	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
250	1		State at this node convergent with a state found outside this clade	Transversion
251	2		State at this node convergent with a state found outside this clade	Transition
285	3		State at this node convergent with a state found outside this clade	Transition
404	2		State at this node convergent with a state found outside this clade	Transversion
466	1		State at this node convergent with a state found outside this clade	Transversion
539	2		State at this node convergent with a state found outside this clade	Transition

Patient 318 – Active disease (case):

Of note in this clade is the predominance of transversions compared to transitions. (9 : 2). Four sites (**229 [residue 77], 301 [residue 101],** 391 [residue 131]**, and 451 [residue 151]**) are first position changes, and three sites are second position changes. Sites in bold represent changes that are not confined to this clade. Four sites (192 [residue 64], 301 [residue 101], 338 [residue 113], and 451 [residue 151]) demonstrated fixation within the clade but variation outside the clade. Again we see a third position 'reversal' at site 390, likely to be indicative of 'wobble', and a first position 'reversal' at site 229 (also a transversion) which may be indicative of a hyper-variable site. Site 229 corresponds to residue 77, which has been shown elsewhere in this study to be of significance.

The transversion : transition ratio also changed to 9 : 2.

Table 8:6 – Patient 318

Site	Position	Within clade	Other clades	Ts/Tv
192	3	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion
229	1		State at this node found within an ancestor, thus representing a reversal	Transversion
236	2		State at this node convergent with a state found outside this clade	Transversion
249	3		State at this node convergent with a state found outside this clade	Transversion
251	2		State at this node convergent with a state found outside this clade	Transition
273	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
301	1	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion
338	2	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion
390	3		State at this node found within an ancestor, thus representing a reversal	Transversion
391	1	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
451	1	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transversion

Within patient analysis

Patient 249 exhibits a high degree of variation *within* the clade, and was therefore investigated in detail.

Within patient 249, there is a predominance of changes (9/19) at position three, likely to indicate genetic drift. There are six changes (**199 [residue 74], 220 [residue 91], 271 [residue 116], 346 [residue 123], 388 [residue 130]** and 460 [residue 154]) at the first position, representing NS changes, and four changes (38 [residue 13], 200 [residue 67], 260 [residue 87] and 500 [residue 167]) at the second position representing possible NS changes. First position sites in bold indicate non-unique changes.

All sites classified as 'reversals' are third position changes, likely to be indicative of 'wobble', with the exception of site 200 [residue 67].
There is a predominance of transitions, with the transversion : transition ratio being 6 : 13.
When the analysis is constricted to cases only, there are eight differences. Sites 15, 111, 199, 231, 260, 321, 346, 500 have all changed their classification. Of these, sites 15, 199, 321, 346, and 500 all changed from being 'uniform outside this clade' to non-unique.

Table 8:7 – Changes within patient 249

Site	Position	Within clade	Other clades	Ts/Tv
9	3	249I and 249M	State at this node convergent with a state found outside this clade	Transition
15	3		State at this node convergent with a state found outside this clade	Transversion
38	2	249I, 249J and 249M	State at this node convergent with a state found outside this clade	Transition
111	3		State at this node convergent with a state found outside this clade	Transition
135	3		State at this node found within an ancestor, thus representing a reversal	Transversion
199	1		State at this node convergent with a state found outside this clade	Transition
200	2		State at this node found within an ancestor, thus representing a reversal	Transversion
220	1		State at this node convergent with a state found outside this clade	Transition
231	3		State at this node found within an ancestor, thus representing a reversal	Transversion
234	3		State at this node found within an ancestor, thus representing a reversal	Transition
260	2		State at this node convergent with a state found outside this clade	Transition
271	1	249L and 249D	State at this node convergent with a state found outside this clade	Transition
321	3		State at this node convergent with a state found outside this clade	Transition
324	3		State at this node found within an ancestor, thus representing a reversal	Transition
330	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
346	1		State at this node convergent with a state found outside this clade	Transversion
388	1	249J and 249C, G, H, I	State at this node convergent with a state found outside this clade	Transversion
460	1	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
500	2	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transition

Patient 029 clusters on the ML tree with patient 413. Therefore, the two were compared as a 'pair'.

Patient 029 - External Branch

Table 8:8 – Patient 029 External Branch

Site	Position	Within clade	Other clades	Ts/Tv
149	2		State at this node convergent with a state found outside this clade	Transversion
192	3		State at this node convergent with a state found outside this clade	Transversion
229	1		State at this node convergent with a state found outside this clade	Transversion
260	2		State at this node convergent with a state found outside this clade	Transition
277	1		State at this node convergent with a state found outside this clade	Transition
278	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
294	3	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition

Patient 029 - Internal Branches

Table 8:9 – Patient 029 Internal Branches

Site	Position	Within clade	Other clades	Ts/Tv
62	2		State at this node convergent with a state found outside this clade	Transversion
126	3	Uniquely derived state, unchanged above	State at this node convergent with a state found outside this clade	Transition

Patient 413 - External Branch

Table 8:10 – Patient 413 External Branch

Site	Position	Within clade	Other clades	Ts/Tv
38	2		State at this node convergent with a state found outside this clade	Transition
96	3		State at this node convergent with a state found outside this clade	Transition
452	2		State at this node convergent with a state found outside this clade	Transition
466	1		State at this node convergent with a state found outside this clade	Transition
543	3		State at this node convergent with a state found outside this clade OR also found at an ancestral state, thus possibly a reversal	Transition

Patient 413 - Internal Branches

Table 8:11 - Patient 413 - Internal Branches

Site	Position	Within clade	Other clades	Ts/Tv
34	1		State at this node found within an ancestor, thus representing a reversal	Transversion
35	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
61	1		State at this node convergent with a state found outside this clade	Transversion
65	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
107	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transition
238	2	Uniquely derived state, unchanged above	Character is uniform outside this clade	Transversion
261	3		State at this node convergent with a state found outside this clade	Transition
283	1	Uniquely derived state, unchanged above	Two or more other states found outside this clade	Transition
467	2	Uniquely derived state, unchanged above	State at this node convergent with a state found outside this clade	Transition

One of the most notable observations from this analysis is the high level of changes (9) found in the internal branches of patient 413 (control) compared to the low level of changes (2) seen in patient 029 (case); this is also paralleled by the conversely low level of changes (5) seen on the external branch leading to patient 413 compared to the relatively high level of changes (7) seen on the external branch leading to patient 029.
Also of interest is the lack of unique sites on the external branch leading to patient 413, compared to a slightly higher level on the external branch leading to patient 029. This trend is then inverted when examining the internal branches, with 029 having no unique sites and 413 having four, with three sites exhibiting fixation within the clade (albeit with variation found in other clades).

Examining patient 029, we see the following amino acids highlighted: *External*: 50, 64, 77, 87, 93, 98; *Internal:* 21, 42. Examining patient 413, we see the following residues highlighted: *External*: 13, 32, 151, 155, 181; *Internal*: 12, 21, 22, 36, 80, 87, 95, 156. Residue 21 is the only similarity between the two patients. There are no similarities between internal and external branches per patient, although there are numerous residues that are proximal. For example, residue 12 is highlighted in patient 413 internally, whilst residue 13 is highlighted externally. The same pattern is found with residues 155 (external) and 156 (internal). Residue 87 is highlighted externally in 029 and internally in 413.

8.12 Amino acid diversity comparison

Previous studies had elucidated known 'hot spots' for variation in the core gene (grey sites, below). Other analyses by us had also elucidated sites that demonstrated variability (red sites, below). Therefore the protein alignment was analysed at each site and the amino acid composition was recorded.

Table 8:12 - Variation found at known hotspots
The table shows the variation found at known hotspots, derived from two separate analyses. Green shading represents differential fixation; Blue shading represents a hotspot found by more than one analysis; Red shading represents a hotspot found in only one analysis; Grey shading represents a hotspot found in the external analysis (Alexopoulou *et al* 2009). Values in circular parentheses indicate the number of patients possessing that amino acid. Notation within square parentheses indicates a single patient with a mixture. Bold denotes a residue not found in the counterpart.

Site	Case (n=6)	Control (n=9)	Epitopes	
12	Ser (6)	Ser (8), [4x Ser**, 5x Asn, 2 x Thr]**		CD4 (HTL) epitope (1-25)
13	Val (5), [9x Val, 3x Ala]	Val (7), Ala (1), **Leu (1)**		
14	Glu (6)	Glu (8), **Gln (1**)		
21	Ser (5), [10x Gly, 2x Ser]	Gly (1), Ser (4), **Ala (1), [2x Asn,** 1x Ser**], [11x Thr,** 1x Ser**] [6x Val, 1x Ala]**	CD8 (CTL) epitope 18-27	
24	Phe (6)	Phe (8), **Tyr (1)**		
26	Ser (6)	Ser (7), **Asn (1)**, **Pro (1)**		
32	Asp (6)	Asp (9)		
35	Ser (5), Ala (1)	Ser (7), Ala (2)		
38	Tyr (6)	Tyr (8), **Phe (1)**		
39	Arg (6)	Arg (9)		
40	Asp (1), Glu (5)	Asp (2), Glu (7)		
50	Pro (5), His (1)	Pro (8), [15x Pro, 3x His]	CD4 (HTL) epitope (50-69) [Ferrari *et al.*, 1991; Jung *et al.*, 1995]	
60	Leu (6)	Leu (9)		
64	Glu (4), **Asp (2)**	Glu (9)		

Site	Case (n=6)	Control (n=9)	Epitope
74	Gly (1), Thr (1), Ser (3), [9x Ser, 3x Gly]	Gly (1), Thr (2), Ser (4), [1x Ser, 2x Gly], [6x Ser, 5x Gly]	
77	Asn (4), Thr (1), [9x Asn, 2x Thr, **1x Ala]**	**Asp (1), Ser (1),** Asn (5), Thr (2)	**B-Cell Epitope** (74 – 89) (anti-HBc / anti-HBe1) [Salfield, 1989]
80	Ala (6)	Ala (8), [1x Ala, **10x Ser]**	
84	Leu (4), **Ala (2)**	Leu (8), **Gln (1)**	
87	Gly (1), Ser (3), Asn (1) [10x Ser, 2x Asn]	Gly (1), Ser (7), Asn (1)	
92	Asn (6)	Asn (9)	
93	Met (5), **Ala (1)**	Met (9)	
101	Leu (5), **Met (1)**	Leu (9)	
105	Ile (6)	**Val (1)**, Ile (6) **[**6x Ile**, 5x Val], [**2x Ile**, 1x Leu]**	
113	Glu (5), **Ala (1)**	Glu (5), [10x Asp, 1x Glu], **[6x His, 2x Gln]**	B-Cell Epitope (107 – 118) (anti-HBc/HBe2) [Celucci et al
114	Thr (6)	Thr (9)	
116	Leu (5), [10x Leu, 3x Ile]	Leu (8), [6x Leu, 2x Ile]	
130	Pro (4), **Ala (1)**, [8x Pro, 4x Thr]	Pro (4), [7x Pro, **4x Ser**, 4x Thr], **Gln (1),** Thr (3)	B-cell epitope (130 – 139) AntiHBc/HBe3 determinan
131	Ala (5), **Pro (1)**	Ala (9)	
146	Thr (6)	Thr (9)	
147	Thr (5), **Cys (1)**	Thr (8), **Ser (1)**	
149	Val (6)	Val (5), **Ile (3),** [**1x Ile,** 2x Val]	
151	Arg (3), Pro (1), **Gly (1), [7x Cys, 5x His]**	Arg (7), Pro (1), **Gln (1)**	
156	Pro (4), **Thr (2)**	Pro (8), **[10x Ser, 1x Phe]**	
174	Arg (6)	Arg (9)	
177	Gln (5), Lys (1)	Gln (9)	

Major IgG binding site (74 – 83)

The above table demonstrates the quasi-species that exist within the dataset. It can be seen that there is high variability of residue composition at certain sites in the core gene, and some have parallels to known immune epitopes.
Of interest in this table is variation found within patients, indicated by square parentheses. Whilst the number of sites that contain variation within patients is similar between cases (8 / 24) and controls (11 / 24), the actual number of patients who demonstrate this variability is higher in the controls (15, compared to 8 in cases). For example, it can be noted that within the controls at position 21 there are three patients who exhibit high levels of variation. This is also seen at positions 74 and 105. It was not documented whether it was the same patients demonstrating 'mixture' for each site or different patients.

Also of interest are the cells highlighted in green, representing 'differential fixation.' For example, it can be seen at site 105 that in the case population, all clones, in all patients, presented with a single amino acid (Isoleucine), compared to the control population, which contained a mixture at this site. This is referred to as 'fixation' - where one amino acid becomes fixed in the clonal population.
Quantification of the level of fixation of a single amino acid at a particular site showed that this is more pronounced in the cases (8) than the controls (4).

Especially of interest in this table however are the six sites (13, 21, 64, 77, and 130) that were elucidated independently by other authors and by this study. Alexopoulou *et al* (2009) showed that these sites are known "hot spots" for variation. Our analysis of diversity highlighted these sites since they contained multiple amino acids. These sites have also been highlighted by analyses run with PAML and MacClade, as well as manual analyses of variation and selection. These sites are also of interest because they are found with known immune epitopes. It can be seen that residues 13 and 21 both are within the HTL epitope (1-25). Residue 64 is found within the 50-69 HTL epitope.

Residue 77 is found within the B-cell epitope, and correlates to the major IgG binding site. Residue 130 correlates to another B-cell epitope (130 – 139), known as the AntiHBc/HBe3 determinant region [Salfeld *et al*., 1989].

Sites highlighted above in red are sites that we have identified independently within our dataset.
It can be seen that there are eleven sites (26, 50, 84, 93, 101, 113, 131, 147, 149, 151, and 156) that were highlighted as 'hotspots' in our dataset that were not found in Alexopoulous dataset. Of these sites, 74% do not correlate to known immune epitopes. Conversely, of the sites identified by Alexopoulou *et al*, 52% do not correlate to known immune epitopes.

8.12.1 HLA Fisher's Exact Tests

A Fisher's exact test was performed on cases versus controls, and restricted to one HLA haplotype per analysis. All sites were examined, however sites that were highlighted by other analyses (e.g. – 13, 130) were examined first and included. Table 8:13 (below) shows that residues 67, 149, 151 and 156 all demonstrate a p-value that approaches significance ($p < 0.15$).

Table 8:13 FET results based on HLA

Residue	A*2402	A*1101	B*5602
13	1	1	1
21	**0.26**	0.39	0.54
26	**0.23**	1	0.49
59	1	1	0.57
67	1	**0.09**	1
74	1	0.99	1
77	1	1	1
79	1	0.39	1
84	1	0.99	1
87	1	1	0.54
91	1	0.39	0.57
97	0.59	0.39	0.54
103	0.48	1	0.51
105	**0.23**	1	0.49
109	**0.23**	1	0.49
113	0.48	0.99	0.51
116	0.38	1	0.39
130	1	1	1
149	**0.10**	1	**0.2**
151	**0.12**	1	0.33
156	**0.12**	1	0.33
180	1	0.99	1