Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Hepatitis B Virus infection in the South Pacific

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

At

Massey University

Palmerston North, New Zealand.

Gabrielle Louise Harrison

1999

This thesis is dedicated to the scent of Sunshine.

Abstract

An exploratory study into the genetics of the hepatitis B virus and its human host in the South Pacific was undertaken to determine direction for future research. This virus is a serious health concern, especially for the indigenous people of this region. The DNA sequence of 14 complete and 2 partial virus genomes were obtained. The viral sequence mutations were investigated and compared with a collected database and current literature. Phylogenetic analysis of the viral sequences was carried out using version 4.64 of PAUP* and SplitsTree. Using the new sensitive method of sequence based typing, HLA-DPA1 allele's were typed in 51 unrelated Trobriand Islanders.

The viral genomes displayed a great deal of variation with many interesting mutations. The results highlight the affect of quasi-species distribution within a host. Phylogenetic analysis identified two hepatitis B genotypes within the South Pacific, HBV-**C** and HBV-**D**. However, the genotype common in northern Europe (HBV-**A**) was not found. The phylogenetic analysis presented a pattern of evolution that resembled that of its human host. The Trobriand Islanders were found to be an extremely homogeneous population, with 86% homogenous for the HLA-DPA1*02022 allele.

The study proved to be very informative, providing the directions of research we aimed for. The Hepatitis B samples demonstrated an interesting pattern of evolution that parallels that of its host supporting a co-evolutionary relationship between host and pathogen, thus hepatitis B appears to be indigenous in the South Pacific. We are presently establishing research to further investigate this pattern by analysing viral samples from Fiji. We have also established research that will investigate the rate of evolution of this virus. The sequenced based typing method proved to be very informative with the ability to detect new alleles. The allele frequency obtained from the Trobriand Islanders agreed with concurrent research and supports the fast-train model of migration into the Pacific. Further work in Fiji will continue with this theme of research as genetic analysis of Fiji has proved to be more complex.

Acknowledgments

David, the next pint in The Turf is on me

Cheers !

My thanks also to Rosalind Murray-M^{c}Intosh, for your guidance and perception.

This thesis would not of happened if it had not been for the kind provision of samples to work with. For this I must thank Prof Wolf Schiefenhovel of the Max Planck-Institute for Behavioural Physiology Von-der-Tann, Andechs Germany, for the Trobriand Islander samples and Dr Robert Hickson for being our liaison. Also thanks to Zlatibor Velickovic of the Wellington Medical School for providing HLA-DPA1 primers and training. Dr Bill Abbott of the Pathology Department of Auckland University and Dr Chris Moyes of the Whakatane Child Health and Hepatitis Foundation provided the hepatitis B virus samples. I would also like to thank Chris Moyes for his many words of wisdom throughout this project and beyond.

To the "plebs", "proliferates", and "boffins", Pete Lockhart who taught me how to do it "his way", Richard Winkworth who taught me how to do it the "right way" and Trish M^{c}Lenachan who taught me how to do it "my way", cheers guys, you did an outstanding job in the face of adversary. Well done! Dan Jeffares and Bennet M^{c}Comish, for 'Bennet's smile', I will always be appreciative, you guys can come check out my mailbox anytime ☺. I am most grateful to Anthony Poole for saving a thousand crashes , Matthew Phillips for his great late night company , Kerryn Slack for being the only person awake when I needed company most; Mark Large, Leon Perrie, and Dave Thomas, for being such good sports (and food for gossip), Carmel Gillman for her serine sanity and Tony Roeven for "world peace". ☺

Mat, Maybee, Mr Pig, Huey, Duey, Leuy (R.I.P.), and George, my menagerie family,

words are no recompense.

Your support, understanding and adventurous soul were, are, and forever will be, my strength.

Thank you.

Contents

List of figures

Chapter one: Introduction

Chapter two: Methods and Software packages

Chapter Three: Results

Appendix C

Appendix D

List of tables

Chapter one : Introduction

Chapter two: Methods and Software packages

Chapter three: Results

Appendix A

Appendix B

List of Abbreviations

%	percent
ALT	alanine aminotransferase
bp	base pairs
ddH2O	deionised and distilled water
DMSO	di-methly sulphoxide
DNA	deoxyribonucleotide acid
dNTP	deoxyribonucleotide triphosphate
E.R.	endoplasmic reticulum
EDTA	ethylenediamine tetraacetic acid
g	grams
HBcAg	hepatitis B core antigen
HBeAg	hepatitis B *e* antigen
HBsAg	hepatitis B surface antigen
HBV	hepatitis B virus
HBV-**A**	hepatitis B virus genotype A
HBV-**B**	hepatitis B virus genotype B
HBV-**C**	hepatitis B virus genotype C
HBV-**D**	hepatitis B virus genotype D
HBV-**E**	hepatitis B virus genotype E
HBV-**F**	hepatitis B virus genotype F
HLA	Human leukocyte antigen
Kb	Kilobase
M	Molar
mg	milligram
MgCl	Magnesium Chloride
MHC	Major Histocompatibility Complex
µl	microliter
ml	milliliter
mM	microMolar
mtDNA	mitochondrial DNA
NaCl	Sodium Chloride
NNB	nearest-neighbour bootstrap
p mol	picomoles

PCR	polymerase chain reaction
RNA	ribonucleotide acid
SDS	Sodium Dodecyl Sulphate
ypb	years before present
°C	degrees Celsius

Chapter One: Introduction

The Haganemnons of Asistus three have the most impatient chromosomes of any life form in the Galaxy. Whereas most races are content to evolve slowly and carefully over thousands of generations, discarding a prehensile toe here, nervously hazarding another nostril there, the Haganemnons would do for Charles Darwin what a squadron of Arcturan stunt apples would have done for Sir Isaac Newton. Their genetic structure, based on the quadruple sterated octo-helix, is so chronically unstable that far from passing their basic shape onto their children, they will quite frequently evolve several times over lunch. But they do this with such reckless abandon, that, if they are sitting at a table, they are unable to reach a coffee spoon, they are liable, without a moments consideration, to mutate into something with far longer arms, but which is probably quite incapable of drinking the coffee. This, not unnaturally, produces a terrible sense of personal insecurity and a jealous resentment all stable life forms or "filthy rotten stinking samelings" as they call them. They justify this by claiming that as they have personally experienced what it is like to be virtually everybody else they can think of, they are in a very good position to appreciate all their worst points. This "appreciation" is usually military in nature, and is carried out with unmitigated savagery from the gun rooms of their horribly beweaponed "chameleoid" death flotilla. Experience has shown that the most effective way of dealing with any Haganemnons you may meet is to run away. Terribly fast !

(The Hitch-Hiker's Guide to the Galaxy; Adams 1978)

1.1 Introduction to the thesis

This thesis is an exploratory study into the genetics of a serious health concern in the South Pacific, the hepatitis B virus (HBV). It is a pilot study to determine directions for future research and the genetics of both the virus and host are investigated. The long term focus of the project is to investigate the genomic sequence of the virus and its phylogeny, as well as the history of the people of the South Pacific and their major histocompatibility complex (MHC) genes.

DNA sequences have become powerful tools in epidemiological studies, for example in Human Immuno-Defficiency disease and Hepatitis C virus (Holmes *et al.* 1995). They enable a greater understanding of the ecology of the diseases. This thesis uses this tool to study HBV in the South Pacific, some aspects of the genetics of both the virus and their human hosts within the Pacific will be investigated. Complete viral genomes from infected Polynesians will be sequenced and analysed, and a start will be made on human leukocyte antigen (HLA) typing, using a new sequence based method.

HBV is the smallest-double stranded DNA virus known to infect humans; it contains multiple overlapping reading frames and replicates through an RNA intermediate by reverse transcriptase. The lack of proof-reading ability of this reverse transcriptase accounts for the high mutational rate of this virus however, the overlapping reading frames place constraints on viable replicants, which restricts the viral evolution. Combined, these factors give HBV a unique mode of evolution which, along with the viral/host evolutionary relationship, remains unresolved (Miller and Robinson 1986; Gojobori *et al.* 1990; Bollyky and Holmes 1998). Co-evolution, cross-species, and recent transfer hypotheses have been suggested but current more sophisticated studies are inconclusive (Mandart *et al.* 1984; Orito *et al.* 1989; Norder *et al.* 1996; Bollyky and Holmes 1998). Clarification of these relationships is of interest, for not only is the molecular biology of the hepatitis B virus unique, but also as it is the major cause of hepatocyte carcinoma world wide, it is a serious health concern.

Infection with HBV can result in a range of clinical states ranging from asymptomatic to fatal. The differences appear to be related to genetics of both the host immune system and the virus (Chisari and Ferrari 1995; Milne *et al.* 1995; Gust 1996). It is estimated that 75% of the world's estimated carriers are within the Western Pacific and South East Asia. Here, where the virus is hyper-endemic (that is, there is a very high proportion of carriers in the population see 1.2.3), patterns of infection vary considerably between villages, cities, countries and ethnic groups (Gust 1996). This observation has relevance when it is noted that archaeological, linguistic and biological studies have shown human migrations through the Near and Far Oceania to be of two separate lineages (Hill and Serjeantson 1989; Kirch 1997; Spriggs 1997). To date little research has been undertaken to clarify the level of genetic admixture between these populations — a relevant issue when considering apparent mode of disease transmission and clearance.

1.2 Literature review

1.2.1 The hepatitis B virus

1.2.1.1 Viral replication

An overview of the replication of HBV is shown in figure 1.1.

The 42nm virion contains an inner core particle which in turn contains a partially double-stranded viral nucleic acid. The minus strand of the viral genome is complete, and bound at the 5' end to the plus strand (which is of variable length). After uptake of the virion into the cell the nucleocapsid migrates to the nucleus where the endogenous viral polymerase completes the plus strand, creating a covalently closed circular (CCC) viral genome of 3.2 kilo bases (Kb). This genome comprises four over-lapping reading frames which produce seven gene products that are expressed by differential transcription initiation. In addition, because of the arrangement of the compact genome with overlapping reading frames (ORF) the control regions are also transcribed and translated into protein (Table 1.1 and Figure 1.2) This overlapping of the sequences regulating gene expression and the gene products themselves, places complex mutational constraints upon the genome (Mason and Seeger 1991; Pugh and Bassendine 1990).

Viral replication results in two types of transcription products, mRNAs for protein synthesis, and pregenomic RNA for viral genomic replication. For protein synthesis, there are four transcripts that encode the core particle/antigen, hepatitis B *e* antigen (HBeAg), surface proteins (HBs_LAg, HBs_MAg, HBs_SAg), viral polymerases (terminal protein, polymerase/reverse transcriptase, ribonuclease H) and the X gene (Table 1.1 and Figure 1.2). The transcripts all terminate at a common site and are produced by the nuclear transcription machinery of the host's cell. The transcripts are transported to the cytoplasm as mRNA where they are translated, after which they are either incorporated into an infectious virion, or alternatively in the case of the precore, core and surface proteins they can be secreted via the endoplasmic reticulum (E.R.) as noninfectious particles (Mason and Seeger 1991; Pugh and Bassendine 1990).

The longer precore transcript codes for the HBeAg, and is a serological marker of viral replication but has no known viral replication function (Chisari and Ferrari 1995). The secreted filamentous surface proteins (HBsAgs) are highly immunogenic epitopes and are present in the blood in excess of the infectious virion (Figure 1.2. and Table 1.1) (Mason and Seeger 1991).

For viral replication, the full genome is copied as RNA and transported to the cytoplasm. The core particle, pregenomic RNA, and the viral polymerase self assemble in the cytoplasm forming the nucleocapsid. Within the nucleocapsid the viral polymerase synthesises the minus strand of the viral genomic DNA and a portion of the plus strand. This nucleocapsid particle migrates into the E.R. where it is enveloped by the surface envelope proteins. The virion is released from the cell without lysis through vesicle-mediated transport (Pugh and Bassendine 1990; Mason and Seeger 1991).

Alternatively, the nucleocapsid particles can disassemble and re-release nascent viral genomes into the cell. These genomes can then re-enter the nucleus and mature to CCC viral DNA, amplifying the pool of transcriptional templates (Pugh and Bassendine 1990; Mason and Seeger 1991).

1.2.1.2 Viral Infection

HBV can be transmitted in a number of ways, which include sexual contact, via blood transfusions and from mother to infant at birth or shortly thereafter (perinatally). The virus can survive for months in an infectious state outside of a host, so it can also be easily spread through contaminated objects, such as razors and tattooing needles. Children infected in early childhood (0-5 years) are generally highly infectious though out primary school years (5-12) as a consequence HBV is quite commonly spread between children via open lesions (Muraskin 1995).

Viral infection may result in a broad spectrum of liver diseases. These range from a subclinical, asymptomatic carrier state with no liver disfunction, (with symptoms perhaps no more than a mild flu, especially for those exposed during infancy and early childhood), to fatal fulminate hepatitis in which death occurs in days. In symptomatic cases, onset of disease is linked with anorexia, vague abdominal discomfort, nausea and vomiting, and often progresses to jaundice. All persistently infected patients are at risk of developing cirrhosis of the liver and hepatocellular carcinoma, particularly those who develop acute liver disease. Five to ten percent of acutely infected adults become persistently infected and develop chronic liver diseases of varying severity. Consequently infection can be disabilitating, severely undermining the quality of life (Chisari and Ferrari 1995; Muraskin 1995; Gust 1996). The prevalence of HBV infection varies geographically. In South East Asia and Oceania the virus is hyper-endemic, with up to 30% of the population infected as asymptomatic carriers. In comparison, in North America and much of Europe about 0.1% of the population are carriers (Gust 1996).

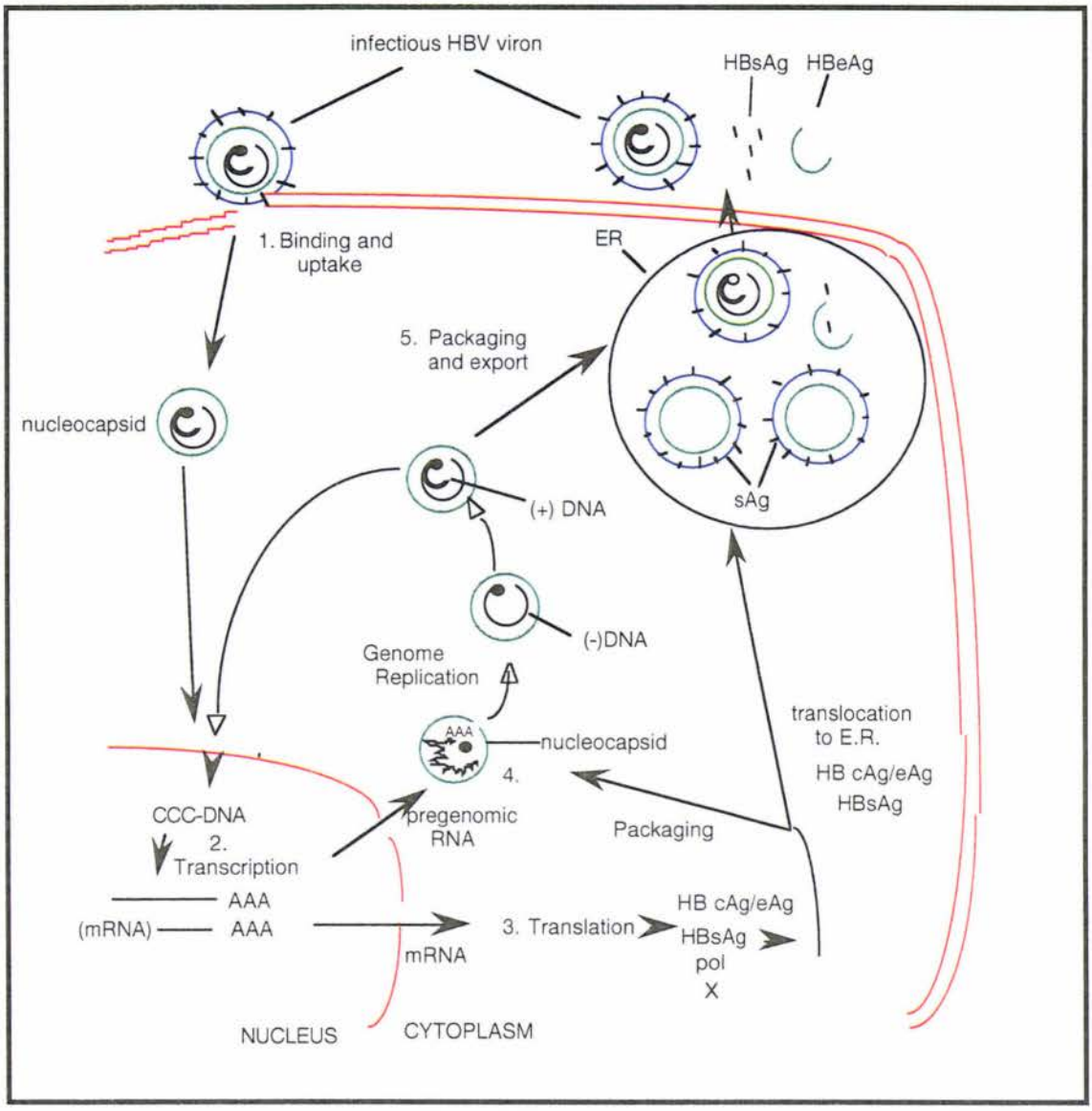


Figure 1.1 Replication cycle of the hepatitis B virus, adapted from Pugh and Bassendine (1990)
The pointed arrows show the direction of replication and the lines indicated what the label represents (see 1.2.1.1 for full description); some of the main steps are:

1. Virus is taken up into the cell and is transported to the nucleus.
2. CCC-DNA transcription produces viral mRNAs.
3. Translation yielding viral structural proteins.
4. Pregenome RNA is packaged into cores with polymerase, and 5´ minus strand protein, enabling genomic replication.
5. Mature cores are packaged into surface antigen particles, which accumulate in ER, and are exported from cell.

Table 1.1 The four genomic transcripts of the hepatitis B virus (Mason and Seeger 1991; Ogata *et al.* 1993; Okamoto *et al.* 1994)

Transcript	Size (bp)*	Overlap	Product	Mapped bases	Function
precore, core mRNA / pregenome	3 500	entire genome	precore protein, (HBeAg)	1816 — 2454	Has antigenic properties, but is not essential for viral replication
			viral DNA pregenome	1903 -1- 1902	Template for viral DNA genome
			core protein	1903 — 2454	Viral genomic packaging/ nucleocapsid
			polymerase	2309 -1- 1625	Polymerase protein
			subunit domains	2309 — 2839	Terminal protein
				2840 -1- 131	Spacer region
				132 — 1163	DNA polymerase/Reverse transcriptase
				1164 — 1625	Ribonuclease H
preS I mRNA	2 400	polymerase X gene	L protein	2850 -1- 837	Large viral surface protein, and has antigenic properties
preS II/S mRNA	2 100	polymerase X gene	M protein S protein	3207 -1- 837 157 — 837	Middle and small Viral surface proteins, both have antigenic properties.
X mRNA	6 50	polymerase precore	X gene	1376 — 1840	Unknown, hypothesized to be a transactive activator

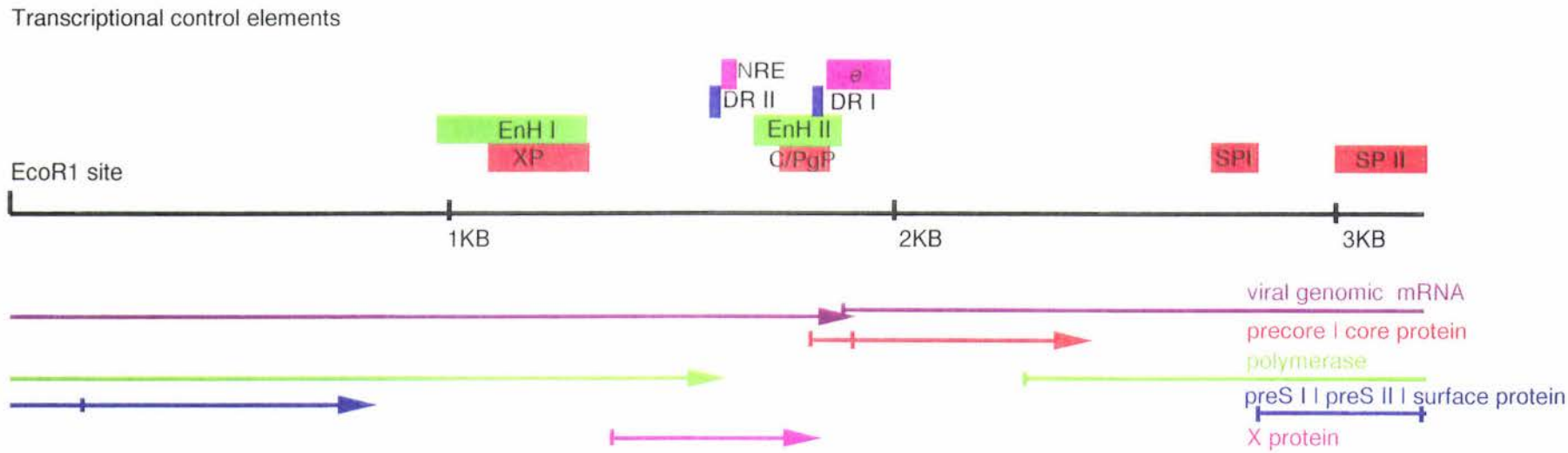
* bp refers to the nucleotide bases more directly, 'base pair'.

Table 1.2 Transcriptional control elements of the hepatitis B virus as described by current literature.

Element	Site	Position (from *Eco* R1 site)	Comments	Reference
preS I promoter region		2711 — 2822	Includes regulatory sequence	Ogata *et al.* (1993)
	HNF	2717 — 2739	Hepatocyte nuclear factor (HNF) binding site	Mason and Seeger (1991)
	TATA box	2784 — 2790		Uchida *et al.* (1995)
preS II/S promoter region		3003 — 3202	Defines the element within a specific region	Ogata *et al.* (1993)
		2967 -1- 370	Defines the element within a broad region	Mason and Seeger (1991)
	NF I	2967 — 2986	Nuclear factor binding site I	Mason and Seeger (1991)
	R/A	3079 — 3134	Sequence activates preS II/S promoter in liver cells	Mason and Seeger (1991)
	E/P	3155 -1- 30	Nonspecific enhancer/promoter like element	Mason and Seeger (1991)
	GRE	341 — 370	Glucorticoid receptor binding site	Mason and Seeger (1991), Ogata *et al.* (1993)
		298 — 369		Uchida *et al.* (1995)
X promoter region		1085 — 1308	Does not include enhancer I (EnH I) site	Uchida *et al.* (1995)
		963 — 1354	Includes EnH I within the X promoter region	Ogata *et al.* (1993)
		882 — 1235	Defines the elements associated with EnH I and the X promoter region	Mason and Seeger (1991)
	UE III	882 — 901	Footprint of undefined protein	Mason and Seeger (1991)
	FP VIII (C/EBP)	977 — 994	Binding site of C/EBP protein CCAAT/enhancer binding protein.	Mason and Seeger (1991)
	FP VII	998 — 1004	NFkB Binding site	Mason and Seeger (1991)
	FP VI	1026 — 1050	C/EBP and HNF I binding site	Mason and Seeger (1991)
	FP IV	1101 — 1119	NF I binding site	Mason and Seeger (1991)
	FP V	1128 — 1139	Binding site of a liver specific protein	Mason and Seeger (1991)
	FP III	1153 — 1172	EFC binding site	Mason and Seeger (1991)
	eH-TF	1153 — 1172	eH-TF binding site overlaps FP II	Mason and Seeger (1991)
	NF I	1159 — 1189	NF I binding site over laps FP II and III	Mason and Seeger (1991)
	FP II	1174 — 1199	C/EBP and AP I binding site	Mason and Seeger (1991)
	FP I	1205 — 1235	NF I binding site	Mason and Seeger (1991)

Table 1.2 Transcriptional control elements of the hepatitis B virus as described by current literature *(continued)*

Element	Site	Position (from *Eco* R1 site)	Comments	Reference
core/pregenomic promoter region.		1744 — 1851		Lopez-Cabrera *et al.* (1990); Okamoto *et al.* (1994)
		1591 — 1850	Includes the enhancer II (EnH II) region	Ogata *et al.* (1993); Uchida *et al.* (1995)
		1599 — 1773	core/pregenomic promoter region and EnH II region	Mason and Seeger (1991)
	FP I	1599 — 1615	Footprint of unknown protein	Mason and Seeger (1991)
	FP II	1619 — 1632	Footprint of unknown protein	Mason and Seeger (1991)
	FP III	1637 — 1703	Liver specific binding sites	Mason and Seeger (1991)
	FP IV	1708 — 1743	Liver specific binding sites	Mason and Seeger (1991)
	FP V	1749 — 1777	Liver specific binding sites	Mason and Seeger (1991)
	FP VII	1833 — 1850	Liver specific binding sites	Mason and Seeger (1991)
enhancer I		966 — 1308	Regulates core/pregenomic promoter site	Ogata *et al.* (1993); Uchida *et al.* (1995)
		1074 — 1234		Dikstein *et al.* (1990) Bichko *et al.* (1985)
enhancer II		1687 — 1805	Regulates surface gene promoter site	Lopez-Cabrera *et al.* (1991); Uchida *et al.* (1995)
		1687 — 1775		Mason and Seeger (1991); Okamoto *et al.* (1994)
		1591 — 1850		Ogata *et al.* (1993)
		1636 — 1741		Bichko *et al.* (1985)
	alpha	1684 — 1733		Okamoto *et al.* (1994); Bichko *et al.* (1985)
	beta	1744 — 1774		Okamoto *et al.* (1994); Bichko *et al.* (1985)
DR1		1826 — 1836		Ogata *et al.* (1993); Okamoto *et al.* (1994)
DR2		1592 — 1602		Ogata *et al.* (1993) Okamoto *et al.* (1994)
e		1853 — 1982	Encapsulation signal of pregenomic RNA.	Ogata *et al.* (1993)
		1849 — 1910		Okamoto *et al.* (1994)
NRE		1613 — 1636	Negative regulatory element of EnH II	Okamoto *et al.* (1994)
		1581 — 1644	CCAC 1617->1620	Park *et al.* 1997



Key	
SPI	preS I promoter region
SPII	preS II/S promoter region
XP	X protein promoter region
C/Pg P	core protein and pregenomic promoter region
EnH I	enhancer I
EnH II	enhancer II
DR I	direct repeat I
DR II	direct repeat II
e	encapsulation signal
NRE	negative regulatory element of EnH II

Figure 1.2 Human hepatitis B virus genomic characteristics demonstrating compact genome with over lapping reading frames; adapted from Mason and Seeger (1991)

1.2.1.3 Vaccine

The first vaccines against HBV were derived from plasma of carriers with the envelope protein in their blood. The individual polypeptides of the surface antigens are immunogenic and when purified are effective antigens for protective immunity. While these plasma vaccines were mass produced on both a local and global scale, in the late 1980s production of recombinant vaccines expressing the HBsAg epitope superseded them (Harford *et al.* 1987; Petre *et al.* 1987; Shiosakil *et al.* 1991; Muraskin 1995).

There are two vaccines available in New Zealand; Engerix B (Smithkline Beecham) and H-B-VAXII (Merck Sharp and Dohme). Both vaccines contain purified surface antigen from either an adw serotype (unknown genotype), or a adw and adr serotype combined (genotype A) (Naumann *et al.* 1993; Saunders 1998) and are produced in culture from a genetically engineered *Saccharomcyes cerevisiae*. These vaccines confer continual immunity by induction of antibodies against the 226 amino acid major antigen determinant S protein of the HBV (Harford *et al.* 1987; Petre *et al.* 1987; Shiosakil *et al.* 1991). These vaccines have been generally successful, but unacceptably high failure rates have occurred in some populations including Polynesians (Kniskern *et al.* 1994). This has prompted much research into methods and timing of vaccinations, such as the use of immunoglobulin and the incorporation of preS II into the vaccine. PreS II + S antigens have been shown to be capable of greater immunogenic induction than S antigen alone (Kniskern *et al.* 1994). These vaccines are presently available but are not in standard commercial use (pers. comm. Chris Moyes, Whakatane child health and hepatitis foundation).

The different status of carrier mothers has a strong bearing on the outcome of child vaccination. Immunisation of babies born to HBeAg positive mothers (those expressing the antigen) is only 50-75% effective in preventing transmission of HBV, whereas it is 95-100% effective in preventing transmission in HBeAg negative mothers (those not expressing the antigen) (Gust 1996). The concern is that perinatal transmitted HBV is rarely cleared, with the child becoming a carrier. In countries where the child-bearing age of mothers is younger the mother is less likely to have seroconverted to HBeAg negative — consequently the child is less likely to respond to the vaccine. This creates a self perpetuating situation, where the female child of a HBeAg positive mother is more likely to become a carrier and pass the infection on to her children.

1.2.1.4 Immunopathogenesis

As previously mentioned, viral infection can result in a range of clinical states. However, all persistently infected patients are additionally at risk of developing cirrhosis and/or cancer of the liver (Chisari and Ferrari 1995). The patient's immune response to the HBV antigens is responsible both for viral clearance and disease pathogenesis during infection. The humoural antibody response to the circulating viral antigens such as the envelope antigens HBs_LAg, HBs_MAg, and Hbs_SAg (see Table 1.1), contributes to the clearance of the infectious particles. The cell-mediated immune response destroys HBV infected hepatocytes through recognition of the HBsAgs, HBV core antigen (HBcAg), and HBeAg epitopes expressed on the cellular membrane. HBV is not considered directly cytopathic to liver cells and most damage is believed to be related to the induced cellular immune response against the viral antigens presented by the HLA molecules on liver cells (Kniskern *et al.* 1994; Chisari and Ferrari 1995).

The distinction between acute and chronic infection is therefore defined by the ability of the individual to react to, and clear, these viral antigens. Vigorous class I- and class II- restricted T cell responses that are polyclonal and multispecific, will usually result in successful clearance, whereas a comparatively weak and narrow response will not. The basis of the difference in response is not known but appears to be related to genetic variations in both the host immune system and the virus (Chisari and Ferrari 1995; Dobson and Carper 1996). The viral genome is susceptible to mutational change, especially within the non-overlapping regions of the precore, core, and the preS I and preS II genes, all of which are immunogens (Miller and Robinson 1986; Pugh and Bassendine 1990; Kniskern *et al.* 1994). Moreover, the HLA immune complex is highly polymorphic (Hill and Serjeantson 1989). It is the interactions of these variables that is believed to result in the variation of different individual clinical states (Pugh and Bassendine 1990; Mason and Seeger 1991; Kniskern *et al.* 1994; Gust 1996).

Many mutations in the viral genome, such as the precore 1896 stop codon mutation, have been extensively investigated (Gunther *et al.* 1992; Ogata *et al.* 1993; Gust 1996;). This specific mutation results in the lack of the expression of the HBeAg (Ogata *et al.* 1993), and a reduction in viral titre in serum. Additional mutations include changes affecting antigenic epitopes of the surface and core genes (Bozkaya *et al.* 1996, 1997). Other studies have also investigated the quasi-species nature of the persistently infecting virus (Alexopoulou *et al.* 1997). Quasi-species are described as a population of viruses that share a common origin, but have distinct genomic sequences as a result of mutation, drift and the impact of selection (Smith *et al.* 1997). HBV infected individuals have been shown to be infected with multiple variants that change over time (Smith *et al.* 1997).

As yet there has been no consistent correlation established between mutations and disease pathogenesis, thus the clinical and epidemiological significance of mutants of the HBV remains to be defined (Gust 1996). The bulk of these studies though, while detailed and informative, have been short term (1-5 years) or have been carried out on recognised infected individuals, limiting comparative analysis of different clinical outcomes.

1.2.2 Hepadnaviridae phylogenetic analysis

To control and prevent infectious diseases a solid understanding of the ecology of the pathogen and its mode of transmission and persistence is a necessary prerequisite. Consequently many molecular epidemiological studies of the phylogenetics of this virus have been carried out, but as yet, the host/pathogen evolutionary relationship remains undefined (Mandart *et al.* 1984; Orito *et al.* 1989; Yang *et al.* 1995). The uncertainty regarding the rate of change and spread of HBV and the relative importance of horizontal transmission between humans, compared with perinatal transmission can lead to quite different understanding of control. Clarification of this relationship and the underlying mechanisms involved has important biomedical applications (Muraskin 1995).

1.2.2.1 Hepadnaviral Evolution

Hepatitis B virus infects primates and is a member of the Hepadnaviridae family. To date three viral subgroups of this family have been identified; avian (duck and heron; DHBV) (Mandart *et al.* 1984), rodent (woodchuck and squirrel; WHV) (Kodama *et al.* 1985), and primate (human, chimpanzee, gibbon, and woolly monkey; HBV) (Norder *et al.* 1996). Human HBV is divided into six genotypes **A-F** (HBV-**A**, to HBV-**F** respectively) based on at least an 8% genomic difference between genotypes (Okamoto *et al.* 1988; Norder *et al.* 1994). There are characteristics intrinsic to each genotype, for example HBV-**A** has a 6 base pair (bp) insertion in the Core gene and HBV-**D** contains a 33 bp deletion not found in the other groups. Geographic distributions, while not absolute, have also been characterised. HBV-**D** is spread world wide and predominates in the Mediterranean area and the near Far East, including India. It has been found in aboriginal populations in Asia, from Indonesia and Papua to Alaska. HBV-**E** to date has only been detected in West Africa. HBV-**F** predominates in the Amerindians, but has also been detected in French Polynesia. HBV-**A** dominates in North-western Europe and Sub-Saharan Africa, while in the indigenous populations of the Far East the HBV-**B** and **C** genotypes are found to be most prevalent. There is also a unique sub-grouping of HBV-**C** found, thus far, only in the Pacific (Courouce-Pauty *et al.* 1983; Onda *et al* 1983; Norder *et al.* 1993).

Phylogenetic studies of this viral family have proposed conflicting hypothesises of viral evolution. These include, an ancient viral/host co-evolution dating the viral divergence to 310 million years before present (ybp) (Chisari and Ferrari 1995), cross-species transfer at about 17-19 million ybp (Norder *et al.* 1996), and independent evolution from about 30 000 ybp (Orito *et al.* 1989), with recent more comprehensive analysis being inconclusive (Bollyky and Holmes 1998).

Co-evolution

Nucleotide sequence comparisons of aligned regions show a 60%-70% homology between human and rodent strains. Between these and the avian strains a lower level of homology is observed, that varies from less than 40 % to 70 % in the more conserved domains. This lower level of homology indicates that the avian virus diverged earlier than the human and rodent strains. This order of divergence resembles that of their host species, and consequently has been used to infer a co-evolutionary relationship between the host and virus. Studies such as that by (Kodama *et al.* 1985), support this hypothesis. They calculated a relative evolutionary distance between the hepatitis viruses based on a ratio of synonoymous (silent) single base changes to the total number base changes in an open reading frame (ORF). While the quantitative comparisons of their results agreed with the above divergence of the hepadnaviridae family, they gave no rate or time of divergence (Figure 1.3a) (Kodama *et al.* 1985). Similarly, Bichko *et al.* (1985) using a rate of $5.1\pm 0.3x10^{-9}$ base changes per site annually (a rate based on the accumulation of silent nucleotide substitutions in DNA based organisms), placed the divergence within the primate HBV, and the primate HBV/WHV divergence from the avian virus, at a times that agreed with their respective hosts' divergence (Bichko *et al.* 1985). The rates of the viral evolution assumed here, in this study may be typical of DNA life forms (organisms and viruses) but I will consider later whether HBV may behave more like an RNA virus and could thus have a very different evolution rate.

Host independent evolution

Orito *et al.* (1989), based on a rate of silent substitutions of 4.57×10^{-5} per site per year proposed that the virus evolved independently of the host, transferring between species. They calculated that the DHBV had diverged 30 000 ybp, the WHBV 10 000 ybp, and the HBV 3 000 ybp (Figure 1.3b). Their work included 18 different hepatitis strains, and a rate of evolution based on the divergence between three complete viral genomes obtained from a single human individual infected perinatally. These divergence times are obviously far more recent than those of the hosts and further the ratios of divergence between the virus species and the hosts do not agree. These results therefore suggests a cross-species transfer (Orito *et al.* 1989).

A strong criticism of this work has been the researchers use of a single individual's HBV infection to represent an entire viral family evolutionary rate, a method that has been repeatedly proved invalid (Bollyky and Holmes 1998; Bozkaya *et al.* 1997). However, Orito *et al.* did conclude their paper by noting the need to refine the evolutionary relationships between and within the Hepadnaviral family.

Cross species transfer

Recently, Norder *et al.* (1996) stated that they had established the presence of natural infection in both gibbon and chimpanzee. These HBV sequences were believed to represent early, distinct lineages within the human HBV strains that diverged later than the earliest diverging human HBV (Figure. 1.3c), implying viral infection predating primate speciation. They hypothesised that perhaps the HBV-**F** genotype represented an early cross-species transfer from a non-ape primate to man, or that an hepadnavirus of a common ancestor to man and ape gave rise to two viral lineages predating the speciation of gibbons from higher apes by 17-19 million ybp. However, the gibbon and chimpanzee sequences have since been discovered to be from recent induced human infection and not natural infection (Bollyky and Holmes 1998).

At present, there is no agreement either on the, rate of evolution of the viral family, nor the extent to which it may transfer between species.

1.2.2.2 Human Hepatitis B Virus evolution.

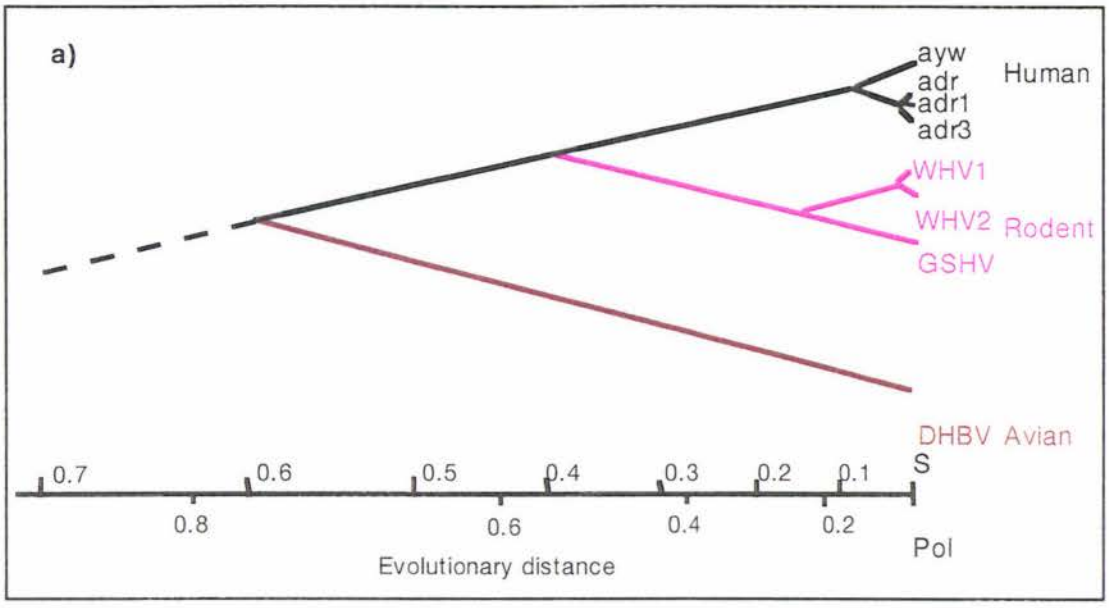
Okamoto *et al.* (1988) began the definition of the separate genotypes which was subsequently continued by Norder *et al.* (1992-1997). Norder's phylogenetic analysis on the complete genomes of these genotypes has placed the first split from the common ancestor with the genotype **F**, and a later split places **B**, **C** and **A** on one side of the tree with **D**, **E**, and the gibbon and chimpanzee strains on the other. Throughout the series of papers though there has been a continual flow of changes; with different analyses using different tree selection criteria and models of evolution giving rise to different divergence patterns of the HBV genotypes (Figure 1.4) (Magnius and Norder 1995; Norder *et al.* 1992-1996).

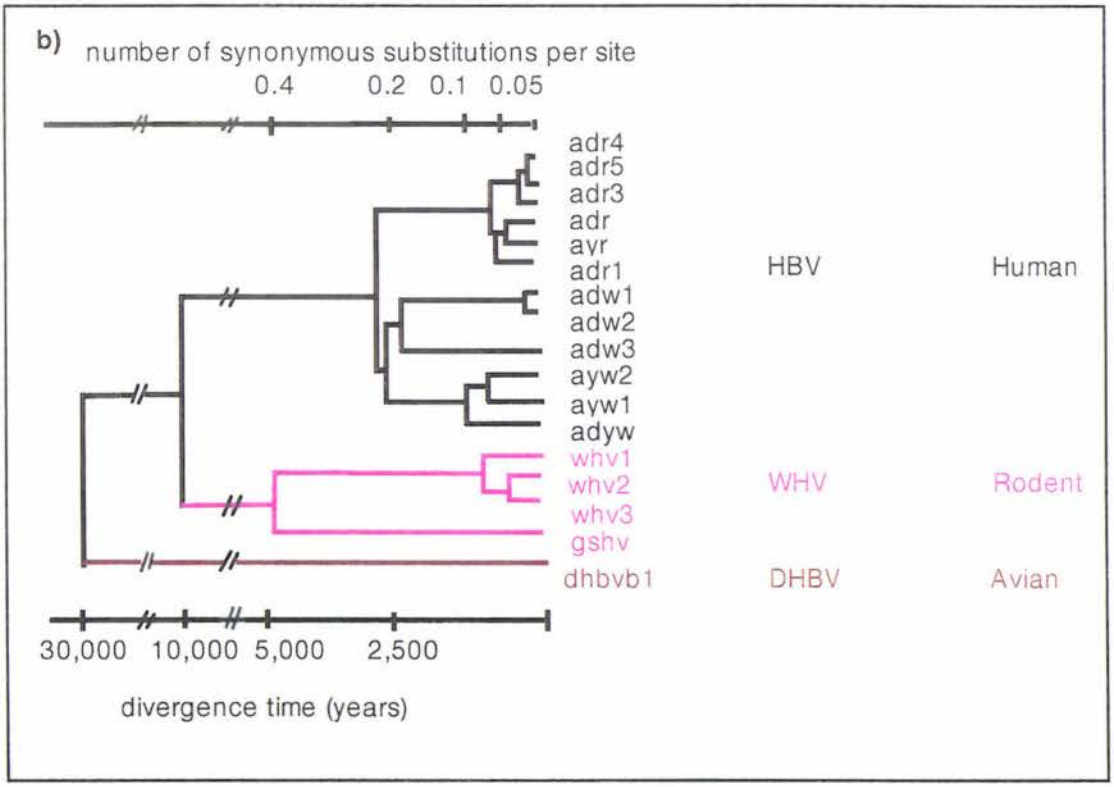
In an attempt to determine a stable, robust phylogenetic relationship of HBV variants, Yang *et al.* (1995) examined the variable substitution rates among sites of the genome. In a two stage approach, models of nucleotide substitution were first compared using a tree that was approximately accurate. Then several candidate trees were evaluated using a working model of nucleotide substitution. The researchers used maximum likelihood, with two models; Felsenstein 1984 (F84) (Felsenstein 1984), and a general reversible process model (REV).

They then combined these with a discrete gamma model to account for rate variation among sites, (in which several equi-probable categories are used to approximate the continuous gamma distribution). To account for the rate difference in the overlapping codon positions they assigned six rate parameters, with all remaining rate variation accommodated by the gamma model. The result was a tree topology that was stable. However, this paper focused on the methods of analysis and did not consider a viral evolution rate estimate, or HBV genotype phylogenetic relationship (Figure 1.5)

Most recently Bollyky and Holmes (1998) set out to reconstruct the complex evolutionary history of the HBV by utilising the complete genomes of 39 mammalian hepadnaviruses and a large sample of small surface gene sequences. They found trees were ambiguous and dependant on the underlying model of DNA evolution. This in turn resulted in more than one human HBV genotype being placed at the branching point of divergence from the last common ancestor HBV-**B** and HBV-**F** (see Figure 1.6) Moreover, in an attempt to assign a rate of nucleotide substitution, multiple core ORF sequences obtained from the same individuals at known intervals were analysed. However, no common rate between two independent data sets was found. They concluded from this that no reliable divergence time or clock could be applied to reconstruct the evolutionary history of this virus.

While it has been established that the avian strains' divergence predates the others, no defined date of divergence nor mode of evolution has been established. Indeed, the phylogenetic evolution of the HBV is still undefined. A number of interesting questions arise out of this work. It is evident that there is considerable uncertainty about the rate of change and spread of HBV, varying from a very recent origin and rapid spread around the world, to an ancient origin predating speciation. HBV-**F** is still considered as the most likely to have split first from the common ancestor, which in turn implies a unique pathogen transfer from the new world to the old world (Bollyky and Holmes 1998).


a)
ayw
adr
adr1
adr3
Human
WHV1
WHV2
GSHV
Rodent
DHBV
Avian
0.7
0.6
0.5
0.4
0.3
0.2
0.1
S
0.8
0.6
0.4
0.2
Pol
Evolutionary distance


b)
number of synonymous substitutions per site
0.4
0.2
0.1
0.05
adr4
adr5
adr3
adr
ayr
adr1
adw1
adw2
adw3
ayw2
ayw1
adyw
whv1
whv2
whv3
gshv
dhbvb1
HBV
WHV
DHBV
Human
Rodent
Avian
30,000
10,000
5,000
2,500
divergence time (years)

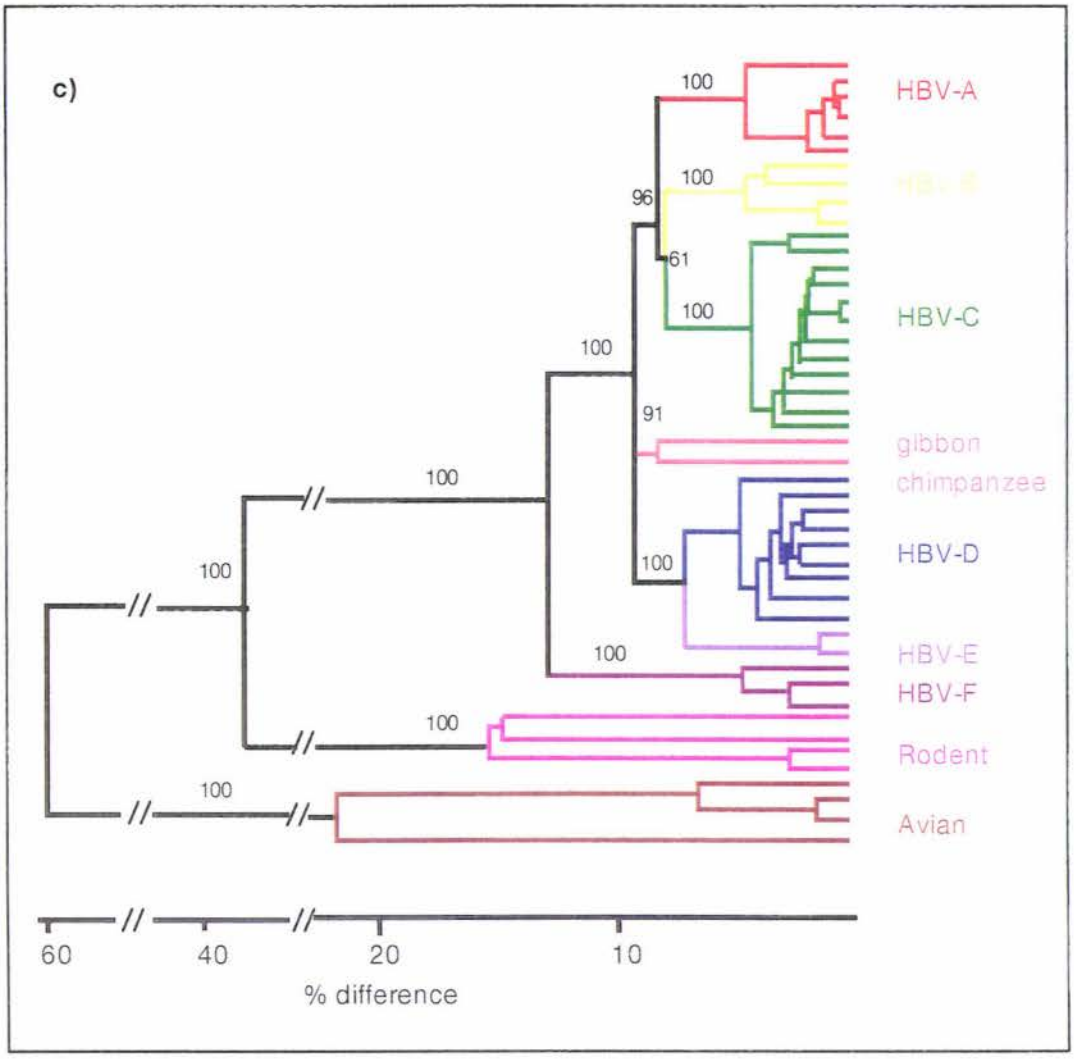
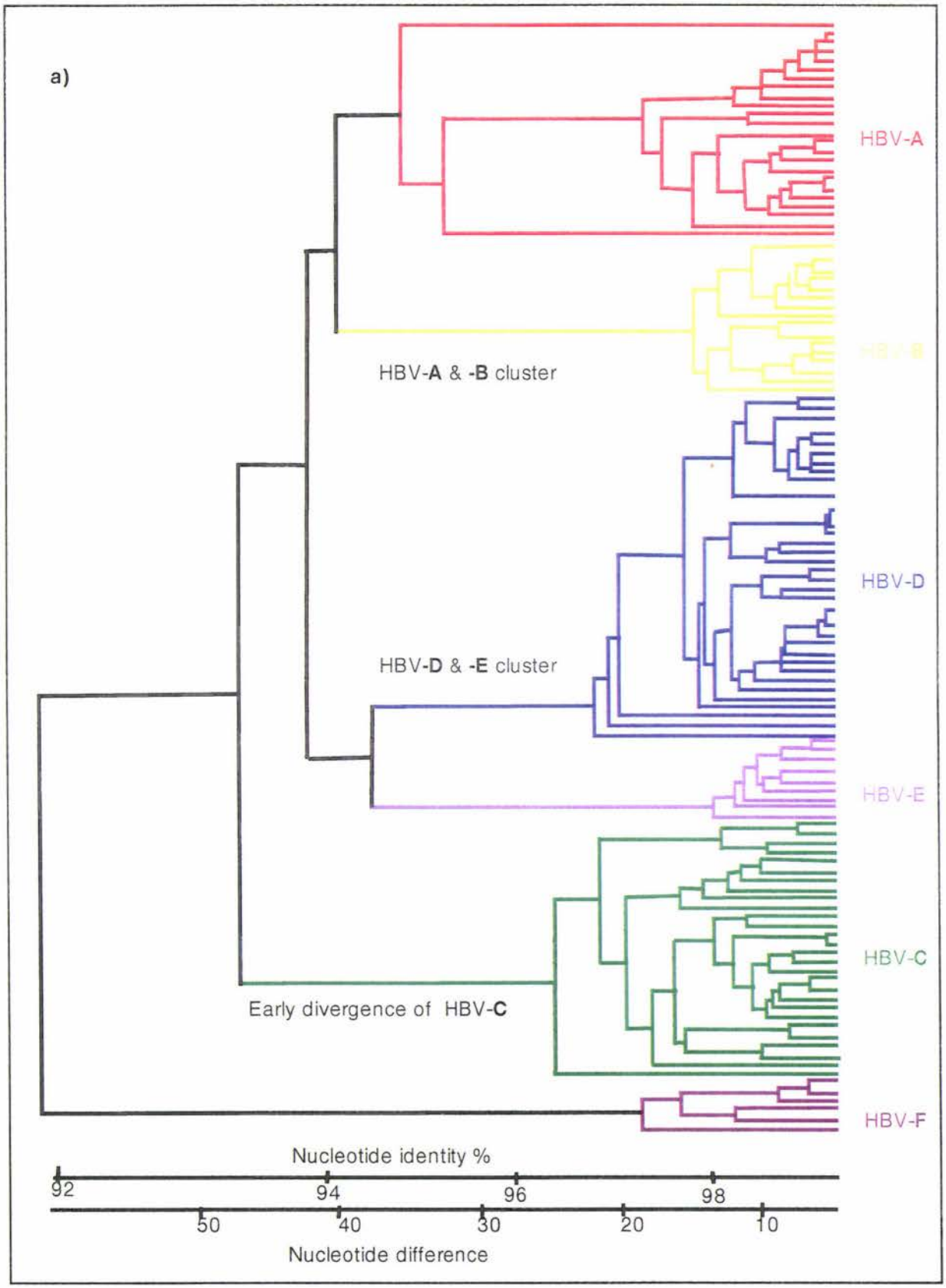


Figure 1.3 Trees representing the phylogenetic relationship of Hepadnaviridae from three separate works using different models. All three trees place the Avian strain as the first to diverge but they vary in the timing of Rodent and Human divergence (see text for details).

Tree **a)** is from Kodama *et al.* (1985). It was modelled on the relative evolutionary distances based on amino acid substitution rates in the polymerase and surface open reading frames.

Tree **b)** is from Orito *et al.* (1989). It was modelled on the relative evolutionary distances based on rate of synonymous substitutions of 4.57×10^{-5} per site per year in the polymerase open reading frame.

Tree **c)** is from Norder *et al.* (1996) It is a UPGMA dendogram, and it demonstrates the gibbon HBV clustering within human HBV genotypes. Boot strap values for this tree are also given.


a)
HBV-A
HBV-B
HBV-A & -B cluster
HBV-D
HBV-D & -E cluster
HBV-E
HBV-C
Early divergence of HBV-C
HBV-F
Nucleotide identity %
92
94
96
98
50
40
30
20
10
Nucleotide difference

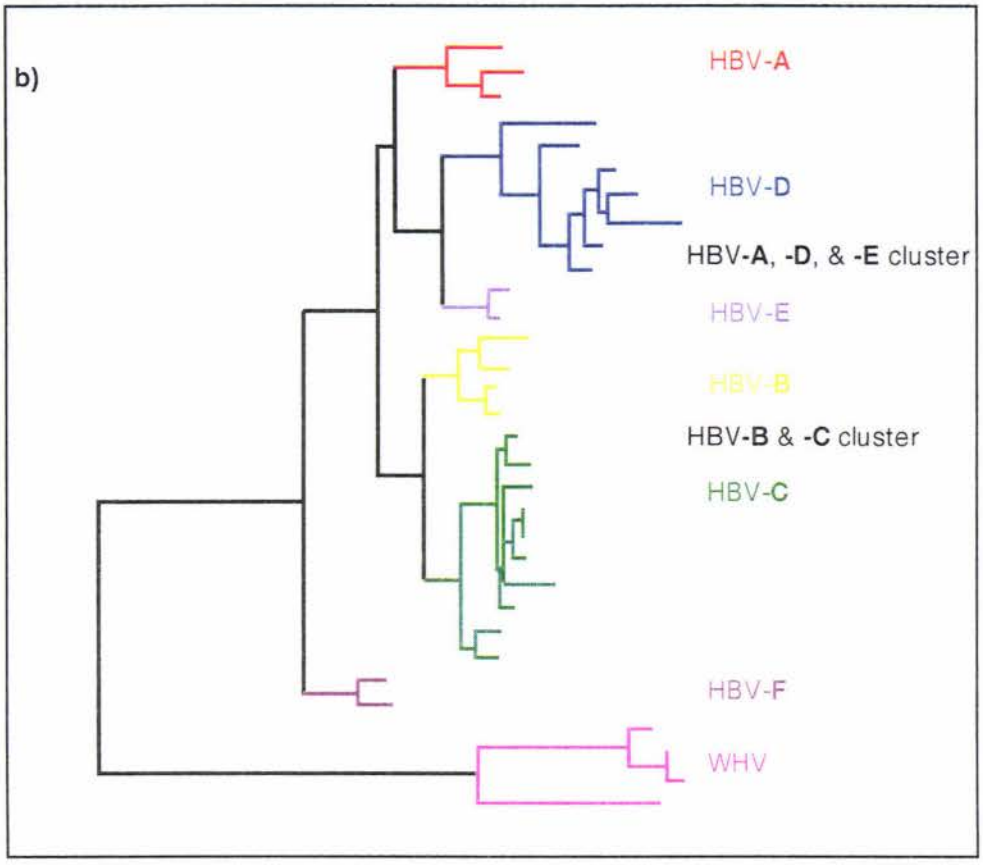


Figure 1.4 Trees demonstrating different divergence patterns of human HBV in the phylogenetic analysis by Norder *et al.* (1992-1996).
Tree **a)** uses UPGMA and implies an early divergence of genotype HBV-**C**, followed by HBV-**A** clustering with HBV-**B**, and HBV-**D** clustering with HBV-**E**. (From Norder *et al.* 1993).
Tree **b)** uses Parsimony analysis. Here genotypes HBV-**A**, HBV-**D** and HBV-**E** cluster together separate from genotypes HBV-**B**, and HBV-**C**. (From Norder *et al.* 1994).

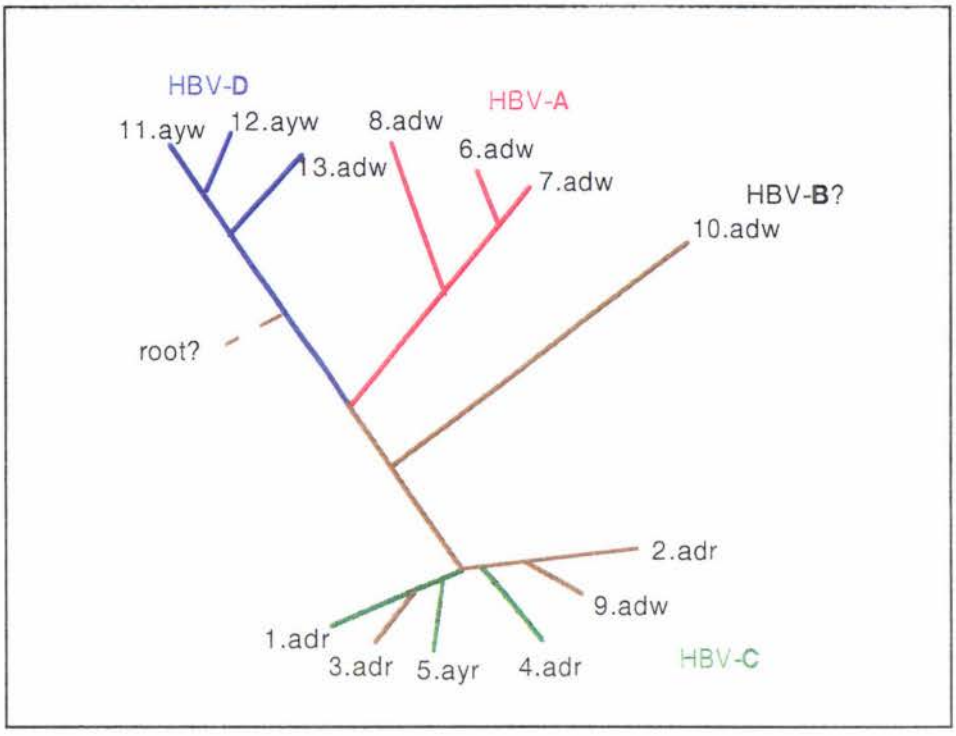
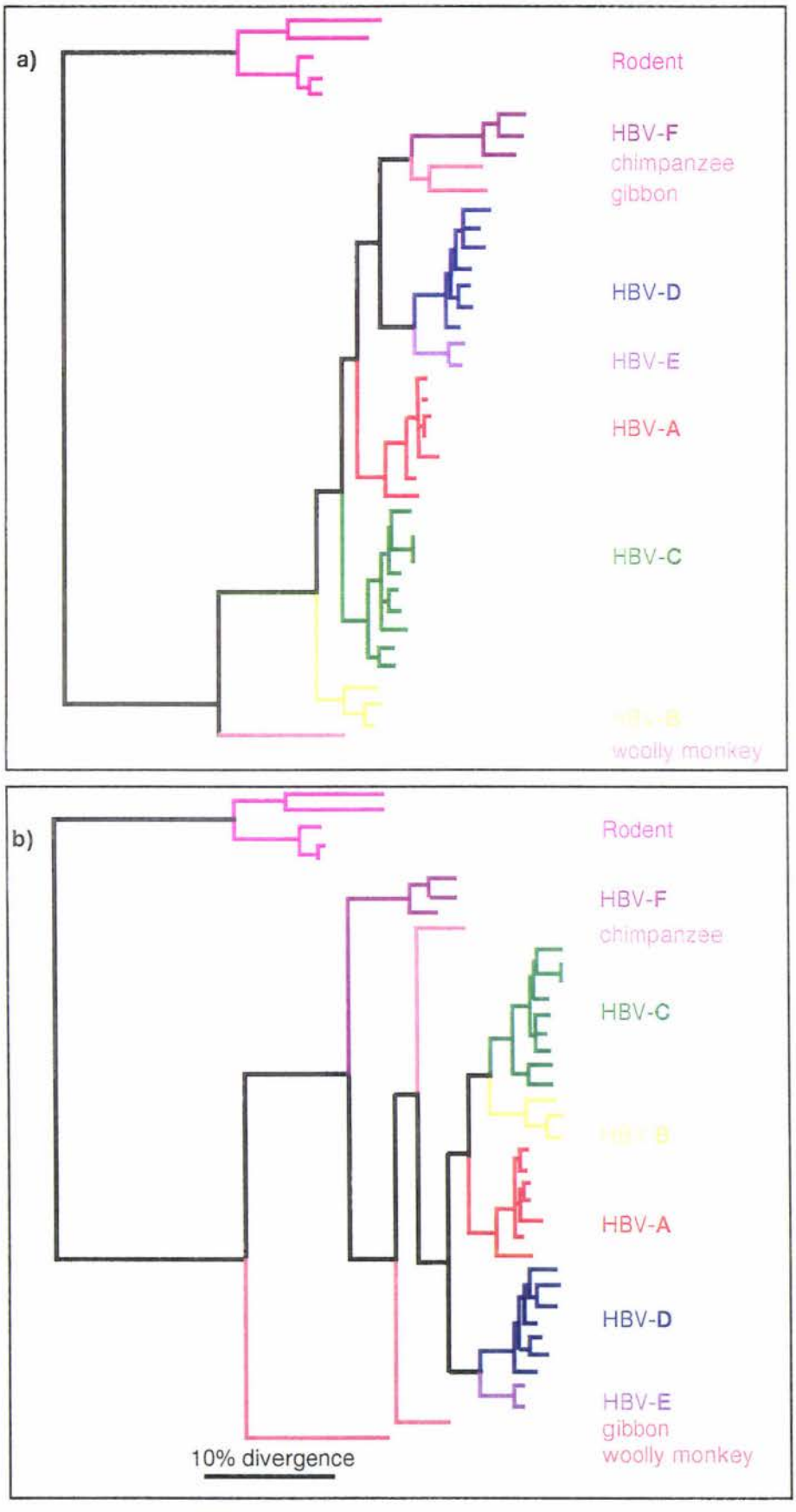


Figure 1.5 Stable tree topology of Human hepatitis B virus (adapted from Yang *et al.* 1995). A stable tree topology was produced from different serotypes, but no time of divergence or rate was calculated, a postulated root position was given and is indicated. Where known, genotypes have been asigned and branches coloured appropriatel y.

Figure 1.6 The ambiguous tree topologies of the hepadnavirus produced by Bollyky and Holmes (1998).
Tree **a)** Uses a gamma distrubution model of maximum likelihood and places genotype HBV-**B** as the more divergent genotype that is, the first to diverge from the last common ancestor.
Tree **b)** Uses the same dataset as tree a) but uses neighbour-joining analysis, here, HBV-**F** is placed as the first to diverge from the last common ancestor.


a)
Rodent
HBV-F
chimpanzee
gibbon
HBV-D
HBV-E
HBV-A
HBV-C
HBV-B
woolly monkey
b)
Rodent
HBV-F
chimpanzee
HBV-C
HBV-B
HBV-A
HBV-D
HBV-E
gibbon
woolly monkey
10% divergence

In summary, these conflicting phylogenetic scenarios lead to quite different understanding of the relative importance of horizontal transmission compared with long term transmission within populations, and consequently clinical control. The question of the underlying mode of host-species evolution has not been answered. Questions addressing the long branch of Genotype HBV-F also have not been considered. While the extensive analysis and data required to answer all these question are beyond the scope of this thesis utilisation of phylogenetic analysis methods still has application in the investigation of the hyper-endemic clinical status of HBV in the Pacific.

1.2.3 Hepatitis B virus and the people of the Pacific

It is estimated that approximately 5% of the worlds population are carriers of the Hepatitis B virus, and of these at least 75% are either of South East Asian or Polynesian ethnicity. In these countries the virus is hyper-endemic (Chisari and Ferrari 1995; Gust 1996).

Much of the information on modes of transmission of HBV is speculative and based on cross-sectional epidemiological studies (Gust 1996). However, most infections occur in these hyper-endemic regions either via perinatal infection or before starting school while the child is part of an extended family that has infected individuals. A lower proportion of transmissions occur in early adult life after onset of sexual activity. Perinatal transmission from carrier mothers who are HBeAg positive to their infants is potentially the greatest concern (Sung 1990). This is because perinatal infection is rarely cleared, with 90% of the infected children becoming chronically infected. Not only does this greatly increase their risk of developing cirrhosis of the liver, but also maintains the cycle of passage between mother and child (see 1.2.1.4) (Sung 1990; Chisari and Ferrari 1995; Gust 1996). The frequency and severity of the disease within the Pacific has led to many regional vaccination programs. These programs have highlighted the different potential vaccination outcome which depend on the carrier mothers HBeAg state. Immunisation of babies born to HBeAg positive mothers is only 50-75% effective in preventing transmission, compared to rates of 95-100% when the mothers are HBeAg negative (Gust 1996). This result appears directly related to the viral titre in the serum, HBeAg positive individuals generally have higher titres compared to HBeAg negative (Gust 1996). Moreover, immunisation programs in countries with high perinatal transmission are only optimally successful when the infant is vaccinated shortly after birth (48hrs). This is only possible where babies are born in hospitals or have immediate contact to health care (Sung 1990; Muraskin 1995).

A fascinating peculiarity of HBV is the variation of the level of infection, which parallels its apparent genotypic distribution. Genotype HBV-**A** is dominant among persistent carriers in North West Europe and the United States, areas which have low HBV prevalence. In the hyper-endemic region of Asia-Pacific, genotypes HBV-**B** and HBV-**C** predominate. This hyper-endemic status is partially explained by the longer duration of the HBeAg-positive replicate state in women of these regions, which highlights the importance of control of perinatal transmission. In sub-Saharan Africa and the Mediterranean, where genotypes **A** and **D** dominate, horizontal transmission is more prevalent (Gust 1996; Arauz-Ruiz *et al.* 1997).

It is postulated that the geographic differences in carrier rates could, to some extent, be explained by variations in the duration of the HBeAg-positive state. It appears that the HBV genotype could be an important determinant for this, and may form the basis for whether perinatal transmission will be common or not (Arauz-Ruiz *et al.* 1997).

In addition, there is variation of infection and related disease within the Asian-Pacific region (Table 1.3) Striking examples include Fiji, where the carrier rate is far higher in the indigenous Melanesian population than in the Indian population. A similar difference is observed in the Maori population compared to the European population in New Zealand. Isolated populations on Pacific islands such as Nauru also show variance from village to village (Gust 1996).

The origin of indigenous peoples of the Pacific is a well researched subject The contemporary belief is that at least two migrations occurred from Asia to the North East. Archaeological, linguistic and biological data all display a clear picture of the human migrations through Melanesia to Micronesia and eventually Polynesia and support the hypothesis of two distinct ancestral populations. Combining this information with geographic and ethnic HBV infection variations may help answer important clinical questions, and may be a powerful tool in clarification of evolutionary structure of the pathogen. Background information on the origin of the Pacific people will now be considered in more detail.

Table 1.3 Prevalence of hepatitis B virus surface antigen in adults within Asia and the Pacific region (adapted from Gust (1996) based on work by J.L. Sung)

Population	Carrier percentage
Taiwan	15
Southern China	12
Singapore	14
Hong Kong	10
Aboriginal (Australian)	5-25
Mediterranean	2-5
Maori (New Zealand)	10-12
Korean	12
Burmese	8-10
Indonesian	5
Indian	5
Japanese	5-15
Anglo-Saxon (Aus/N.Z.)	0.1

1.2.4 The people of the Pacific

The earliest know human habitation of Suhul (the combined continent of Australia, Papua New Guinea (PNG) and Tasmania) occurred 50 000-60 000 (Roberts *et al.* 1990) ybp in Northern Australia. A date of 30 000-40 000 ybp is believed to be the time of human arrival in PNG (Spriggs 1997). The accepted migration path is via the Suhul connection to Sunda-land (Asia, Borneo, and Java) (Stevens 1985). Archaeological evidence reveals a second great migration into Melanesia beginning 5 000 ybp, that spread into many Melanesian Islands, before crossing across thousands of kilometres of open ocean into remote Oceania, and stretching out to New Zealand, the Easter Islands, and Hawaii. These second people are believed to be of South East Asian origin (Kirch 1997). Thus, through the combined disciplines of archaeology, linguistic and genetics, the peoples of the Pacific have been recognised as coming from two distinct populations (Bellwood 1987; Gibbons 1994; Kirch 1997; Spriggs 1997). Nevertheless, clarification is lacking as to the level of genetic association between these two populations. This information is important for not only will it give a more precise view of the peoples' history, but also, it will aid in the further understanding the varying patterns of HBV infection in the Pacific. Here the comparative analysis of allele frequencies in the HLA complex has the potential to provide information directly addressing both these questions. HLA allele frequencies give an autosomal measure of genetic admixture between populations and, as these genes are part of the immune complex, comparative analysis will aid in the interpretation of the varying patterns of HBV infection. While the HLA complex has been investigated extensively, it has not been done so in Pacific Islanders at the DNA sequence level.

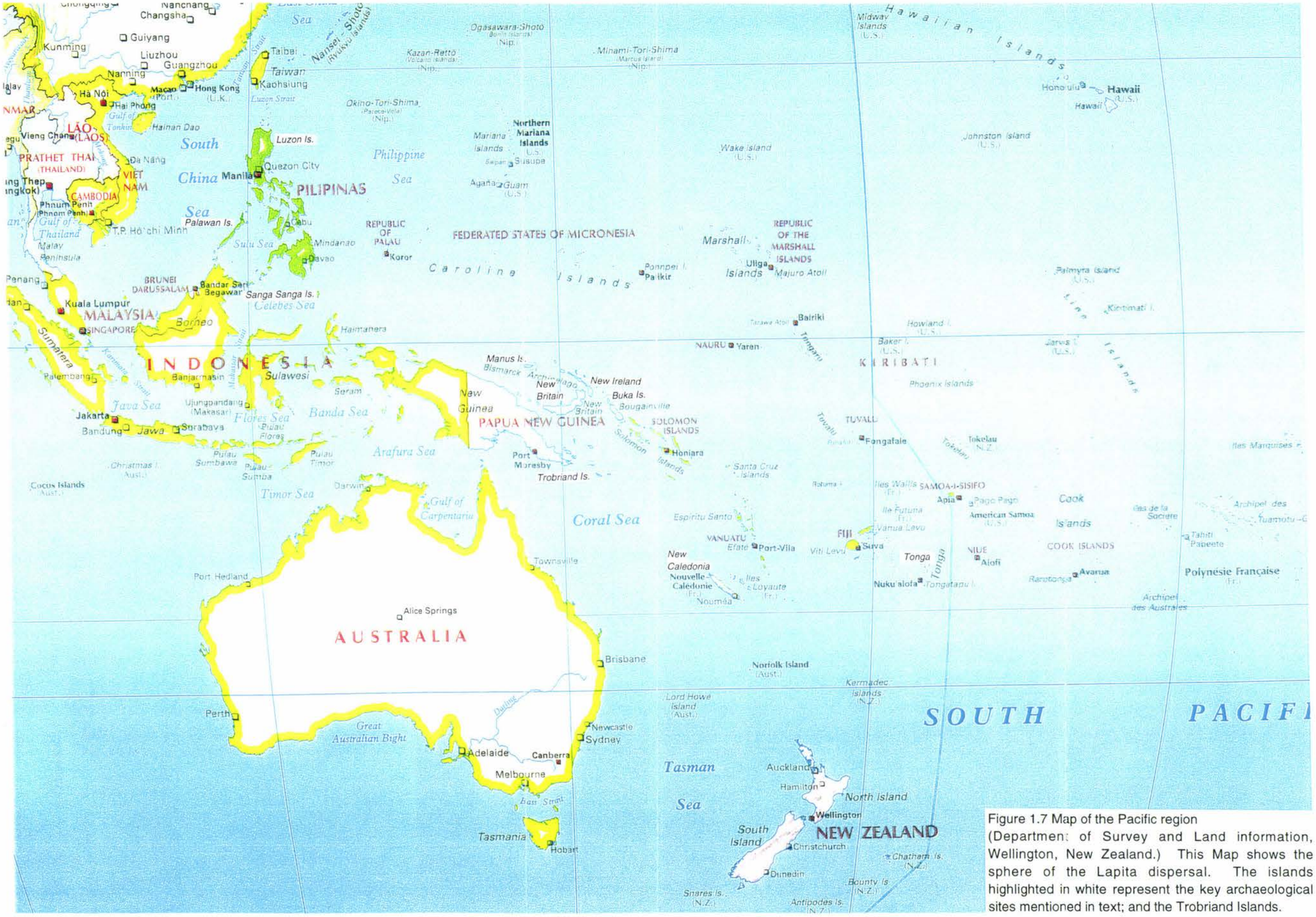


Figure 1.7 Map of the Pacific region
(Department of Survey and Land information, Wellington, New Zealand.) This Map shows the sphere of the Lapita dispersal. The islands highlighted in white represent the key archaeological sites mentioned in text; and the Trobriand Islands.

1.2.4.1 Archaeological evidence

Archaeological comparison of sites that date within the Pleistocene (from 60 000 - 10 000 ybp) and the mid Holocene (10 000 - 3 000 ybp) periods has defined and differentiated two Pacific populations.

Pre-Lapita

While there are only a few sites, there is clear evidence that the Asian peoples of Sunda-land, journeyed into the Melanesian Islands during the mid to late Pleistocene 50 000 - 20 000 ybp (Kirch 1997; Spriggs 1997). The sites from within the earlier Pleistocene era contain sea shell, fish and animal bone deposits, an array of steep edged flakes and other undistinguished stone tools. Excavation of the Kilu site of Buka Island (see Figure 1.7) revealed a focus on plants, but not an established agricultural economy (Spriggs 1997).

From 30 000-6 000 ypb there was a noticeable change in the raw tools used to produce stone tools however, the stone tools themselves changed little, and so minimal advancement in technology occurred. Analysis has shown that processing of root crops also stayed the same within that period (Spriggs 1997).

In general, the sites through out the Pleistocene and early Holocene periods show continual marine exploitation and multiple introductions of species from mainland New Guinea. Examples include bandicoot species *(Echymipera),* at about 13 000 ybp and the possum, *(Phalanger orientals),* about 16 000 ybp to New Ireland. Crop plants were also introduced; *Canarium,* an important food source, is found in most of the sites through out the islands. Obsidian trade between New Britain and New Ireland is dated to 20 000 ybp (see Figure 1.7). Transformations therefore occurred upto 5 000 ybp, that clearly indicated boating technology and minimal advancement in technology (Spriggs 1997).

Lapita

The Lapita legacy is fundamentally an archaeological construct. It is named after the 'Lapita' excavation site on the western coast of New Caledonia where a large collection of potsherds was recovered that matched earlier discoveries in Tonga and Watom. The legacy is defined by its distinctive decorative pottery, which dates within the mid- to late- Holocene (5 000 - 3 000 ypb). This Lapita culture is considered ancestral to the Polynesian people. Like a trail of bread crumbs, it dots its way across the Pacific, leaving a distinctive path of decorative pottery, stretching from Watom and New Caledonia to Fiji and Tonga, and spanning the geographies that divide Polynesia and Melanesia (Spriggs 1997).

Archaeological assemblies predating Lapita have been found in Taiwan where the Ta-p'en-k'eng culture is typified by fine "red-ware pottery" dating between 6 500 and 6 000 years ybp (Bellwood 1987; Kirch 1997). In addition, associated with the Lapita culture are the presence of polishing stones, pitted cobbles, and fish hooks of bone and shell. Similar Neolithic assembles have been found in the Luzon Islands of the Philippine archipelagos about 4500 ybp, and the Palawan and Sanga Sanga Islands off the coast of Borneo (Kirch 1997). The designs fade and lack elaboration as they move from West to East out of the Melanesian Islands, but are not found in any similar form predating 5 000 ybp in the Melanesian islands (Kirch 1997; Spriggs 1997).

The Lapita people travelled with pigs, dogs, chickens (Kirch 1997; Spriggs 1997), the Pacific food rat (Matisso-Smith *et al.* 1998), and lizards (Austin 1999) and thus demonstrated animal husbandry skills. The Lapita people established villages with their houses on stilts, and avoided sites where malaria may have been rife. Compared to sites pre dating 5 000 ybp, ornamental and tool artefacts show marked advances in sophistication, including stone and shell adzes, and marine shell artefacts such as bracelets, necklaces, knives, and fish hooks. Extensive trade of obsidian from Manus occurred throughout the Bismarck Archipelagos, Solomon Islands, and into Vanuatu. Obsidian from New Britain, was traded from Borneo to Fiji post Lapita arrival. Previously, trade had been limited to the island grouping associated with the mining area (Kirch 1997; Spriggs 1997).

Thus, archaeologically there was a sudden influx of relatively advanced technology and economic activity between 5 000 and 4 000 ybp, which can be traced back to Asia. As well as animals introduced of Asian descent. The stilted houses, and the red slipped pottery can be traced back to Taiwan.

1.2.4.2 Linguistic Evidence

Linguistics, unlike archaeology, uses contemporary patterns of variation. The variation encodes and reflects a history, in which daily life, culture and society can be reconstructed. This is achieved using lexical reconstructions which take words (lexemes) and their probable meanings (semantic values) and create a proto-language (Kirch 1997). These lexical reconstructions can refer to known archaeological objects, for example, within the Proto-Oceanic language (see later) '*Kuron*' means pottery vessel. Other examples use common technological ability; for example, the peoples of the Pacific were known sailors, and in the Proto-Oceanic language words such as '*layaR*' refer to sails, and '*saman*' for outriggers. These lexemes highlight known common culture (Bellwood 1991; Kirch 1997).

In the 1930s, linguist Otto Dempwolf drew on a host of lexicons, grammars and dictionaries, and firmly established the Austronesian language family. It was this work which first recognised the Polynesian languages as a subgroup of the Austronesian language family (Spriggs 1997). Since then much analysis has been carried out developing a widely accepted phylogeny, and placing a probable root of the Austronesian languages in the Formosan languages spoken by indigenous Taiwanese. The Formosan languages spoken by the indigenous Taiwanese form some of the first order groupings of the Proto-Austronesian languages. Another group of Formosan languages forms the first order group of Malayo-Polynesian (MP) languages, which subdivide on a West to East geographical axis. Of the Western MP, Central MP and Eastern MP languages, it is the Eastern MP Language that precede the Oceanic languages (Figure 1.8).

All the languages spoken by Polynesians are Oceanic Languages and fall within the Lapita dispersal (Bellwood 1991; Kirch 1997). The implication of all the Austronesian languages deriving from a single ancestral language is that effectively one Austronesian colonisation of the Pacific occurred (Kirch 1997). The Oceanic languages are dispersed in a bush-like manner with three high-order clusters, suggesting that the primary dispersal of the Oceanic language was in Eastern Melanesia. Further, this bush-like model depicting the dialects implies that there was a rapid dispersal of the peoples into the central Pacific, while still maintaining communication networks over a considerable distance (Kirch 1997).

However, there are non-Austronesian languages in the Pacific, namely the Papuan languages of Australia and the Highlands of PNG. These languages do not show the same cohesion as the Austronesian, with at least 12 distinct Papuan language groups. The variability of the Papuan languages is attributed to their age (at least 40 000 years), and the physical barriers of mountainous terrain blocking social interactions. (Bellwood 1991; Kirch 1997).

The correlations between the archaeological and the linguistic data are obvious; both clearly demonstrate two distinct populations within the Pacific. The first population was less mobile, with agricultural skills, and were the ancestors of the Highland New Guineans. The second population was highly mobile with a much broader spatial distribution and range of movement, these Lapita people were the Polynesian ancestors. Linguistics clearly defines the two populations and their movement within the Pacific (excluding Australia). Archaeology puts a time reference to both migrations, the first between 60 000 and 30 000 ybp, and the second from South East Asia, between 5 000 and 4 000 ybp. However, there are still discrepancies over the finer details; did the Melanesians come with the Lapita people? After the Lapita people? Or at all to Polynesia? What sort of unification was there? Biological data is required to answer these questions.

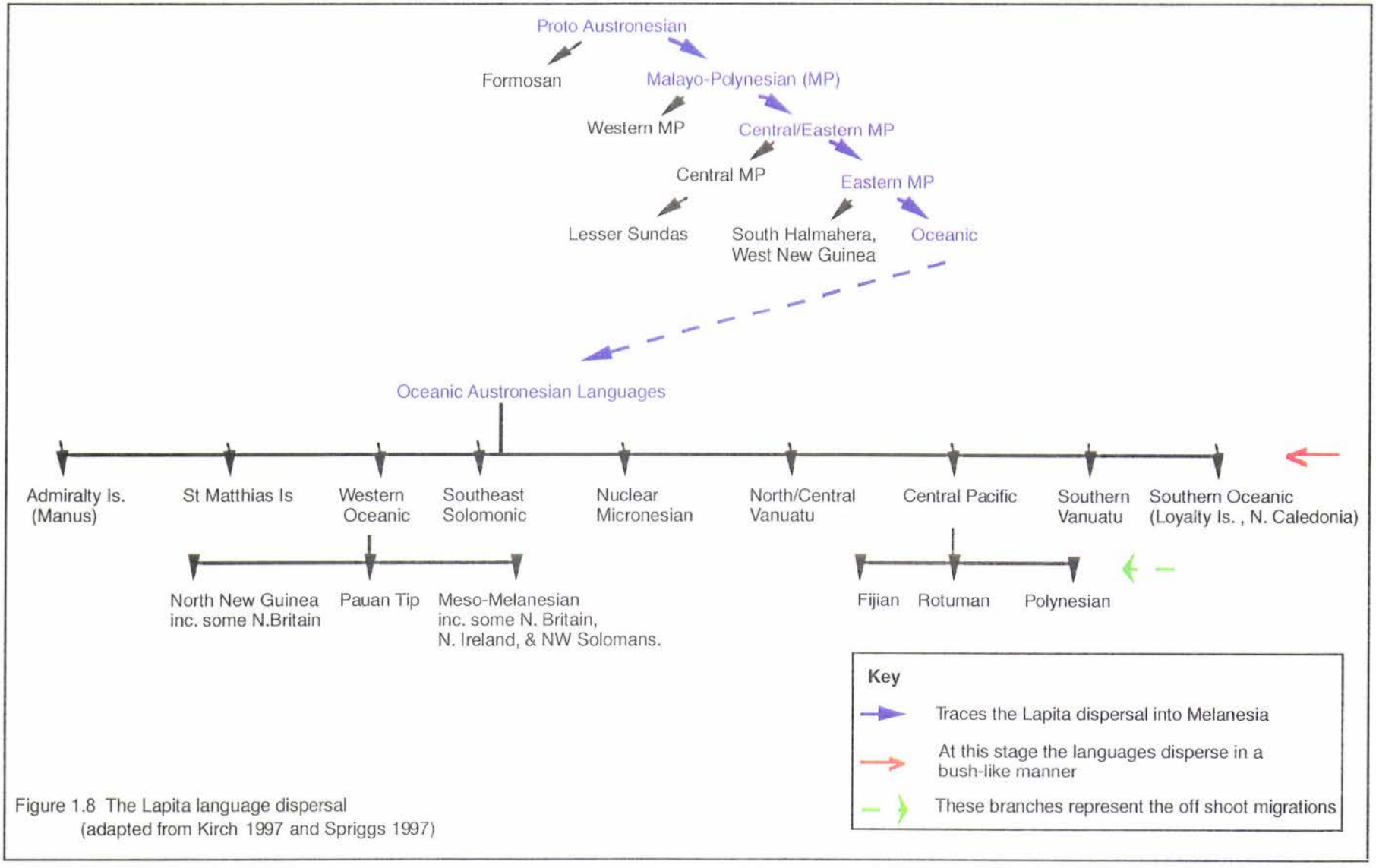


Figure 1.8 The Lapita language dispersal
(adapted from Kirch 1997 and Spriggs 1997)

1.2.4.3 Genetic evidence

The genetic data available for population genetics is continually expanding and becoming more informative (Hill and Serjeantson 1989; Lum *et al.* 1994; Mairias *et al.* 1994; Nagy *et al.* 1996). Though often not initially identified from a population biology perspective, there is a vast amount of data available which is informative. Traditional markers such as red cell antigens, serum protein groups and complement components, as well as immunoglobulin allotypes, have been extensively studied on a global scale (Hill and Serjeantson 1989). These markers exist in a polymorphic state and their variant frequencies have revealed similarities and/or difference between population groups. However, the data has often been obtained from a non-population biological perspective and has been surpassed by contemporary markers from the mitochondrial genome, the HLA complex, microsatellites and more recently Y-chromosome markers (Lum *et al.* 1994; Mairias *et al.* 1994; Nagy *et al.* 1996).

Mitochondrial data.

The human mitochondrial genome is a 16.5 kb circular molecule that encodes polypeptides essential to the mitochondrial (mt) energy generating pathway, together with rRNA and t-RNAs necessary for mt protein synthesis. mtDNA exhibits greater heterogeneity than nuclear DNA, and shows little evidence of recombination (but see Hegalberg *et al.* 1999a). Due to a lack of "proof reading" or repair enzymes, and apparent relaxed constraints on the translational apparatus giving a greater tolerance of mutations, the genome evolves by simple sequential mutations, at a rate of up to ten times that of nuclear DNA. Inherited from the maternal line in a haploid manner, the mt genome provides an excellent tool for reconstructing the maternal aspects of ancient human migrations (Hill and Serjeantson 1989; Merriweather *et al.* 1991; Lorenz and Smith 1994; Wallace 1995).

Restriction fragment length polymorphism (RFLP), sequence specific oligonucleotide probes (SSO), and direct sequencing have all been used to measure the variation within the mt genome (Hill and Serjeantson 1989; Lum *et al.* 1994).

The most informative mt marker of the colonisation of the Pacific to date has been a 9 bp deletion found by Cann *et al.* (1983) in the V region of the mt DNA (Murray-MnIntosh *et al.* 1998). Typically the V region contains two tandem repeats of 9 bp, the deletion of one repeat occurs in East Asian populations and is found in 93% of Polynesians (Harihara *et al.* 1992; Hertzberg *et al.* 1989; Red *et al.* 1995 Soodyall *et al.* 1996; Sykes *et al.* 1995). Of the 150 individuals sampled by Herztberg *et al.* (1989) 100% of the Samoans, Maoris and Niueans, 82% of Fijians, 8% of Tolais, and 14% of Coastal New Guineans had the deletion.

However, it was found to be absent in Highland New Guineans and Australians. Lum *et al.* (1994) demonstrated three deep maternal lineage clusters within the Pacific which appeared to have a common ancestor of about 85 000 ybp.

The group one lineage contained the 9 bp deletion and usually three transition substitutions in the control region. Common in remote Oceania, group one included 95% of the Native Hawaiian, 90% of the Samoan, and 100% of the Tongan individuals sampled. Group one also contains non-Polynesian individuals such as Indonesians, Native Americans, Micronesians, Malaysians, Japanese and Chinese (Lum *et al.* 1994).

The second group includes Polynesians who do not have the deletion and share four distinct single base substitutions. This second group occurs at a frequency of less than 10% in the Polynesian populations sampled, and was hypothesised as an ancestral maternal lineage group of Papuan Melanesians (Lum *et al.* 1994).

The third lineage group has not been found in Hawaii, and appears to link Samoa to Indonesia. Overall the conclusions reached by Lum *et al.* (1994) was that Lapita people mixed with the existing Melanesian peoples and thus carried forward an assortment of maternal genotypes derived from two distinct geographic sources to the remote Pacific Islands.

Y Chromosome Information

The Y-chromosome describes the male history in a fashion that is complementary to the mt DNA. The paternally-inherited Y chromosome is partially haploid and this section lacks recombination, so variations on this haploid section can be used readily for population genetic studies. The rate of mutation varies at these sites from $5x10^{-7}$ mutations per locus per generation for base substitutions, to a few percent for the minisatellites. Therefore the genetic information contained with the Y-chromosome has the potential to clarify and further explain the human migration across the Pacific (Hurles *et al.* 1998).

In many cases the mtDNA data does not agree completely with that of the autosomal HLA genes, the obvious assumption being that males and females were differentially represented during migrations. Fiji is one such example (Hurles *et al.* 1998) Population markers have recently been described for the Y chromosome on the non-recombining section results are patchy so far within the Pacific. A strong European influence has been found, highlighting the sensitivity of the Y chromosome to recent admixture events (Mairias *et al.* 1994; Kayser *et al.* 1997; Hurles *et al.* 1998).

Microsatellite markers specific to Polynesians have been found that suggest a large inflow of proto-Polynesian male lineages into island and coastal Melanesia as well as Polynesia however, their high mutation rate however makes it difficult to trace events older than 8 000 ybp (Hagelberg *et al.* 1999). Y chromosome data has been slower to accumulate than mt DNA data but it is showing potential for future research Hagelberg *et al.* 1999a, 1999b).

Major Histocompatibilty Complex

The initial interest in the major histocompatibility complex (MHC) came from its immune response role, eliciting tissue rejection during organ transplants, and the associated autoimmune diseases, (such as insulin dependant diabetes) (Hill and Serjeantson 1989; Serjeantson 1989; Riley and Olerup 1992;). However, the characters that created the interest — the large number of alleles per locus, the presence of population specific variants, and linkage disequilibrium (the tendency of HLA alleles to be inherited in clusters) are the key features that make this complex a powerful tool in the genetic analysis of populations (Hill and Serjeantson 1989; Riley and Olerup 1992). The MHC is a large complex with many genes. The most informative MHC genes to population genetic studies are those of the class I and class II type. This complex has a multitude of alleles per gene which have the potential to provide more than 10 000 billion genetically different variants. A few alleles have been found to be unique to certain regions and some alleles appear together in individuals more often than would be predicted on the basis of population gene frequencies and linkage disequilibrium (Hill and Serjeantson 1989). While the extreme diversity can hamper population affinity analysis between broad groupings, in confined areas like Oceania, the unique alleles and the linkage disequilibrium have proved to be very informative (Gao *et al.* 1992; Nagy *et al.* 1996; Velickovic *et al.* 1998; Hagelberg *et al.* 1999a, 1999b).

The alleles of this complex have been studied via serological, cellular, biochemical and DNA analysis. Studies using RFLP of the Class II alleles and the linkage disequilibrium relationships between the Class I HLA A, B, and C, and Class II HLA DR, DP, DQ revealed heterogeneity between the Pacific populations and is extensively reviewed in Hill and Serjeantson's "The Colonisation of the Pacific". What was described in this review and subsequent studies (Gao *et al.* 1992; Hagelberg *et al.* 1999a; Velickovic *et al.* 1998) can be summarised in the following points.

- The migrating populations were small.

- Micronesians have had an independent source of HLA genes that are not found elsewhere in the Pacific.

- The basic similarity between Australia and the Highland PNGs implies a common ancestry. The comparative lack of some HLA alleles in the PNG peoples suggests smaller founding populations.

- PNG coastal and island populations are similar, sharing a sub-stratum of HLA features found in Australians and Highland PNGs. This supports the proto-Polynesians co-existing with the Melanesians, as they moved into Polynesia.

This uniquely polymorphic MHC/HLA complex has proven to have the potential to provide an autosomal measure of admixture between populations via comparative analysis of allele frequencies (Gao *et al.* 1992; Hill and Serjeantson 1989; Hagelberg *et al.* 1999a). However, to differentiate between many of the HLA alleles, traditional techniques like MLC, serology, SSO, RFLP and amplified fragment length polymorphism (AFLP) have proved to be inadequate, especially in identifying new variants in a population. With the advances in the polymerase chain reaction (PCR), fully automated sequencing and heterozygote resolution through dye-primer chemistry, direct sequencing of the complex's alleles has become possible and has the potential to identify new alleles (Gao *et al.* 1992; Versluis *et al.* 1993; Velickovic *et al.* 1998). While the HLA complex has been investigated extensively, this new sequence based typing method is just beginning to be applied to the populations of the Pacific (Nagy *et al.* 1996; Velickovic *et al.* 1998).

This study used the MHC HLA-DPA1 gene in its analysis, as it has only 13 different known alleles and does not required examination of multiple exons for typing. Other genes are morepolymorphic but were beyond the exploratory scope of this project.

1.3 Aim

To control and prevent infectious diseases, a solid understanding of the ecology of the pathogen, transmission and persistence is necessary. To determine such factors knowledge of the size, spatial distribution of the host population, the movement of infected and susceptible hosts and the nutritional status of the host population is crucial (Dobson and Carper 1996). In the case of hepatitis B viral infection, this knowledge is even more paramount due to the unacceptably high failure rate of the vaccine in the hyperendemic populations of the Pacific. Thus this thesis is an exploratory study investigating hepatitis B virus infection within the Pacific by sequencing the complete viral genomes of infected Polynesians. This sequence data will be correlated with clinical states and carriers status of the donor individuals, and the underlying evolution of the viral strains will be examined. To further elucidate the underlying ecology of the virus, HLA-DPA1 typing of a selection of Pacific individuals was undertaken to combine with other such data to further probe the Pacific population and spatial admixture.

Chapter Two: **Methods and Software packages**

The aim of this thesis is to investigate some aspects of both the genetics of HBV infection and the human host within the Pacific,

- by sequencing the complete genomes of a selection of HBVs, and
- by HLA-DPA typing of a selection of Pacific individuals.

Samples of HBV infected New Zealand Maoris, Tongans, Samoans and Niueans were provided by Dr Chris Moyes, of the child health and hepatitis foundation, Whakatane, and Dr William Abbott, Pathology Department, Auckland University Medical School, in the form of 600 μl of serum. The Whakatane samples were of asymptomatic carriers detected in a HBV survey of Kawarau in 1985, and who have since been monitored by the foundation. The Auckland samples were from Pacific Island women living in Auckland who were HBV asymptomatic but were found to be HBV positive in routine testing during pregnancy. In addition to this, 51 blood samples of unrelated Trobriand Islanders were provided by Prof Wolf Schiefenhovel, Max-Planck-Institute for Behavioural Physiology Von-der-Tann, Andechs, Germany, for HLA-DPA1 gene analysis. The Trobriand Islanders were ideal for this study because they are geographically positioned within the flow of the Lapita dispersal and have had minimal interaction with Europeans or other Pacific Islanders (Figure 1.7) (Nagy *et al.* 1996; Hagelberg *et al.* 1999a, 1999b).

With both types of samples, DNA was extracted and amplified using PCR. All the HLA-DPA1 amplifications and most of the HBV genome amplifications were sequenced directly from the PCR product by automatic sequencing. However, in some cases a few sections of the HBV genomes required cloning. As much relevant information as possible was collected, such as, carrier status, liver function, sex, and ethnicity (for HBV serum providers). A database of viral sequences was collected from Genbank for combined analysis, as was HLA allele data within the Pacific. A combination of clinical and phylogenetic analysis was applied to the data obtained. The methods used are described below.

2.1 DNA Extraction

HBV samples

Viral DNA was isolated using protocols outlined in Gunther and Meisel (1992). 100 μl of serum was incubated at 65°C for four hours with 20 mM Tris/HCL, 10 mM EDTA, 0.1% SDS and 800 $\mu g/ml$ of Gibco BRL Proteinase K in a total volume of 200 μl. DNA was extracted with equal volumes of phenol and then chloroform (200 μl each) and precipitated overnight in 2.5 volumes of absolute ethanol with 20 μg Boehringer Mannheim t-RNA and 0.1 volumes of 3M Sodium Acetate, pH 5.3. This was followed by a 25 minute 4°C spin at 13 000 rpm and then a 500 μl, 70% ethanol spin wash. The dried pellet was re-suspended in 10 μl H_20.

Trobriand samples

The Trobriand blood samples were provided as a 300 *µl* TE buffer DNA extract. These were prepared from 2 *ml* of whole blood using the method of Kan *et al.* (1977). This entailed Proteinase K and RNase digestion, followed by three phenol/chloroform extractions, and precipitated in two rounds of absolute ethanol. After the second wash the dried DNA was brought up to volume in TE buffer.

Table 2.1 Materials, solutions and cell strain and genotype used.
Except where otherwise stated in the text all chemicals were AnalaR grade, from BDH supplied by BioLab Scientific LTD and media, growth factors, and antibiotics where from Gibco BRL, Life Technologies.

DH5α™ competent cells:
Gentoype F-Φ80dLacZΔM15(LacZYA-argF) Cl 169 cleoR recA1 and A1 hsdr17(r_k, m^+_k) phoA supE44λ-thi-1 gyrA96 rela1 (Life Technologies MAX efficiency).

Gel Loading buffer (10X):
0.44%(w/v) Bromophenol blue, 0.44%(w/v) xylene cyanol, 27.5%(w/v)Ficoll 400.

LB Broth :
10 grams Bacto®-tryptone, 5 grams Batco®-Yeast Extract, and 5 grams NaCl per litre.

LB -IPTG/X-Gal/Ampplates:
LB Broth plates containing 0.5 mM IPTG, 80 *µg/ml* X-Gal and 100 *µg/ml* ampicillin.

Lysozyme solution:
10 *mg/ml* Boehringer Mannheim lysozyme, Tris Cl pH 8.

RNAse solution:
2 *mg/ml* of Sigma RNAse A type 1A R4875, 10mM TrisCl pH 7.5 boiled for 15 minutes to make DNAse free.

STET buffer:
0.1M NaCl, 10 mM Tris-HCL, 1 mM EDTA, 5% Triton X-100.

TAE Buffer: Tris Acetate EDTA buffer.
89mM Tris HCL, pH8.0, 1mM EDTA.

TE Buffer: Tris EDTA buffer.
10 mM Tris HCL, pH 8, 1 mM EDTA,.

2.2 PCR Amplification

PCR is a robust, efficient means of amplifying DNA sequences. It is based on a procedure of separating a double stranded DNA template using a high temperature of 94-96°C, followed by a rapid drop to 40-60°C for 30-60 seconds, to enable oligonucleotide primers to anneal to the single stranded template. The annealing temperatures are based on the melting temperatures of the individual primer pair and homogeneity to the template; higher annealing temperatures allow more specific amplifications.

Following primer annealing, the temperature is raised to 74°C, where the polymerase enzyme replicates the DNA template. Each repeat of these cycles amplifies the DNA template in a multiplicative manner. Due to the sensitivity of this reaction with only a few nanograms of DNA required, contamination is a real concern. Consequently, many controls were used to create "DNA sterile conditions" and to monitor contamination. Aerosol resistant tips were used to stop DNA transfer and "blank" control reactions containing no intentionally added DNA were run, to check that the reagents were clean (Saiki *et al.* 1988; Palumbi *et al.* 1991). All reactions were done in the 0.2 *ml* MJ Research PCR tubes in a MJ research PTC-200 Peltier Thermal Cycler using a heated lid, with simplified hot starts in which reaction tubes were not added until the thermal cycler was up to initial temperature. BioLab Scientific Qiagen PCR Kits were used

2.2.1 HBV samples

Primers

The primers were based on primer positions from Horikita *et al.* (1994), Uchida T. *et al.* (1995), and Asahina *et al.* (1996a). They were designed using 24 entire HBV sequences acquired from Genbank which were subsequently aligned using Clustal X (Thompson *et al.* 1994). A consensus sequence was created from this alignment using Applied Biosystems Inc. (ABI) MT Navigator PPC version 1.0.2.b3 software. This consensus sequence was then used in Oligo®4.03 primer design software to design primer combinations (Table 2.2). While all primers worked for PCR reactions, four did not give sequencing products, therefore new primers were designed based on the Pacific sample sequences obtained, these are HB3R2, HB4F2, HB4R2, and HB6R2.

PCR Amplification

PCR amplifications of the HBV DNA were in 20 *µl* volumes containing 1X Qiagen PCR Reaction Buffer, 1X Qiagen PCR Q-Solution, 250 µM of each dNTP, 0.5 µM of each primer, 1µl of DNA extract, and 1 unit of Qiagen *Taq* DNA polymerase. All the reactions used an initial step of 94°C for 120 seconds and a template denaturation step of 94°C for 60 seconds, completed by a final elongation step of 72°C for 4 minutes. However, times and temperatures for the *Taq* polymerase extension and primer annealing steps varied depending on the required template amplification length and primers respectively (Table 2.1). The amount of serum limited the amount of viral DNA available. Finding constant sites large enough to map strong primers to was also difficult, due to the variable nature of the HBV genome. Consequently, to get a useable amount of amplified template, 'nested reactions' were often required. Nested PCR involved an initial standard PCR amplification, and quantification of the amplified product by visualisation on an agarose gel. Up to 5 ng of the first round product was then re-amplified using either; primers that bind within the DNA template; or one primer that was used in the first amplification with the other binding internally. If excess DNA was added then non-specific copying could occur which results in an unusable product (Palumbi et al. 1991). For both the first round, and nested reactions, 40 amplification cycles were used.

Table 2.2 Hepatitis B virus primers that were used and designed in this study

a) ABI *nucleotide degenerate code*

Code	Nucleotides
M	A + C
K	T + G
R	A + G
S	C + G
W	A + T
Y	C + T
B	T + C + G
D	A + T + G
H	A + T + C
N	A + T + C + G
V	A + C + G

A negative control containing no DNA was used in each set of reactions to check for contamination within the reagents. Where appropriate, positive controls with HBV DNA known to amplify were also used. Aerosol resistant tips were used at all times to prevent any DNA contamination.

Table 2.2 Hepatitis B virus primers that were designed and used in this study (continued).

b) Name, sequence, melting temperature, and genomic position of the primers.

Primer	Sequence (5'—3')[A]	Tm° C[B]	Genomic position [C]
HB1F	CCKCCTCCTGCYTCCACCATTC	70	3121
HB1R	ACAGCAACADGAGGGAWRCWTAGAG	72	547
HB2F	TTCCTCTKCATCCTGCTGCTATG	68	400
HB2R	CAGCAAANCCYAAAAGWCCCACAA	68	997
HB3F	TGAGTCCCTTTWTRCCKCTRTT	60	780
HB3R1	YGGWACSGCAGAYRRAGAAG	58	1492
HB3R2	CGTGCAGAGGTGAAGCGAAG	64	1612
HB4F1	GGTCTGGAGCRAWVMTYATC	56	1307
HB4F2	CGATCCATACTGCGGAACTC	62	1324
HB4R1	AGCTTGGAGGCKTGAAMAGT	56	1859
HB4R2	AATTTATGCCTACAGCCTCCTA	62	1774
HB5F	CTTGGACTYYCWGSRATGTC	55	1655
HB5R	GTKGATARGATAGGGGCATTTG	62	2304
HB6F	ACWGCACTCAGGCAAGMWRTT	56	2057
HB6R1	ATTAACTRSMTDHTCWGGATAAT	60	2701
HB6R2	TTWGGATAAAACCTAGCAGGCAT	60	2667
HB7F	ATGTTAGTAYYCCTTGGACTC	58	2446
HB7R	GGGCTCCACCCCAAMAKDC	58	3065

A Degenerates are based on the ABI degenerate code

B T_m values are the melting temperatures based on the GC method, where A/T = 2°C + G/C = 4°C

C Positions are numbered from the *Eco*R1 site or equivalent

c) Primer combinations and the annealing temperature and extension times used.

Primer combination	Annealing temp °C	Extension time (seconds)
HB1F HB2R	57	70
HB1F HB1R *	57	60
HB2F HB2R *	57	60
HB3F HB5R	50	120
HB3F HB4R(1&2) *	52	90
HB5F HB5R *	50	60
HB6F HB8R	50	90
HB6F HB7R *	50	70
HB7F HB8R *	52	60

* nested primer combinations

2.2.2 Trobriand samples

Primers.

The primers were provided by Zlatibor Velickovic, Department of Pathology, Wellington Medical School. Designed using Oligo® 4.03, they match constant sites of the intron 1 and intron 2 which surrounds the DPA exon 2, and they amplify a fragment of 483 bp. In addition, M13 tails were added to enable the use of standardised M13 Dye primer chemistry sequencing kits.

DPA exon two primer sequence

Upstream M13 tailed primer 5' TGT AAA ACG ACG GCC AGT GCT CCT TCT TCT TCC CCA TA 3'

Downstream M13 tailed primer 5' CAG GAA ACA GCT ATG ACC CCCTCT CAT CCC TTC CAG TT 3'

PCR Amplification

PCR reaction mixtures of 20 μl volumes contained 1X Qiagen PCR Reaction buffer, 3.75 mM $MgCl_2$, 250 μM of each dNTP, 0.5 μM of each primer, 1 μl of DNA extract, 1X Qiagen PCR Q-Solution, and 2.5 units of Qiagen *Taq* DNA polymerase made up to volume with deionised and distilled (dd) H_2O. The reactions began at 94°C for 120 seconds for complete DNA strand separation, and were followed by thirty cycles of: 94°C for 30 seconds for standard cycle template separation; 63°C for 30 seconds, to allow primer binding, and finally 72°C for 30 seconds to allow *Taq* polymerase extension. To complete the copying, the final 72°C was maintained for 4 minutes. The amplified products were maintained at 4°C until they were prepared for sequencing.

2.2.3 Confirmation of PCR product

Each PCR amplification, including all controls, was visually checked for product. Ten percent of the reaction volume was loaded onto 1% TAE agarose gels and electrophoresed at 100V for 40 to 60 minutes. GIBCO BRL 1 kb or 1 kb-plus ladders were run beside the PCR product to check the product was the correct size. The gels were stained in ethidium bromide and photographed over UV light for DNA visualisation.

2.3 PCR template purification

Due to their interference with the sequencing reaction, it was necessary to remove all substrates from the PCR product. Two different BioLab Scientific Ltd, Qiagen Kits were used, subject to requirements.

2.3.1 Qiagen QIAquick™ PCR purification kit protocol

The Qiagen columns contain a silica-gel membrane which is designed to purify single or double stranded PCR products ranging from 100 bp to 10 kb from free dNTPs, polymerases, and salts that would inhibit sequencing reactions. The method used was as recommended by the Qiagen QIAquick™ handbook. Five volumes of buffer PB is added to 1 volume of the PCR reaction, this was then bound to the silica membrane via centrifugation for 60 seconds at 13 000 rpm where the PB buffer containing the PCR solution was passed through the QIAquick™ spin column into a collection tube. The flowthrough was discarded and the column was washed with 0.75 *ml* PE buffer and centrifuged for 60 seconds, again the flowthrough was discarded and the column re-spun for 60 seconds to dry the column. To elute the DNA bound to the membrane 30 *µl* of Qiagen elution buffer was dispensed to the centre of the column, and was is left to stand for 60 sec, before elution via centrifugation into a clean 1.5 *ml* microcentrifuge tube. All spins were done at 13 000 rpm.

2.3.2 Qiagen QIAquick ™ Gel extraction purification kit protocol

This protocol is designed to extract and purify DNA of 100 bp to 10 kb from standard low melting point agarose gels in TAE or TBE buffer. At times secondary bands appeared due to excess template and/or nonspecific primer binding, and under these conditions a method of gel purification was used. This entailed loading the entire PCR amplification sample into a 1% TAE agarose gel and separating the bands by electrophoresis. Under sterile conditions, the correct fragment was excised from the agarose gel and dissolved in three volumes of Qiagen buffer QX1 to 1 volume of gel (e.g. 100 *mg* of gel requires 300 *µl* QX1 buffer) at 50°C. Maintaining the correct pH of 7.5 was important for binding the DNA to the column membrane, this was monitored by the colour of the QX1 buffer which is pH sensitive. Once dissolved the equivalence of one gel volume of isopropanol was added to the sample mix. This was then added to the QIAquick™ spin column in the 2 *ml* collection tube and spun for 60 sec. The flowthrough was discarded, and a wash of 0.75 *ml* of buffer PE was centrifuged through the QIAquick™ column for 60 seconds. The flowthrough was again discarded and columns were re spun to remove any residual ethanol from the column. To elute the DNA, 30 *µl* of Qiagen elution buffer was added to the column. This was left to stand for one minute, before being spun again for one minute to elute a purified template into a clean microcentrifuge tub. All spins were at 13 000 rpm

2.3.3 Quantification of DNA

Automatic sequencing reactions require the DNA template concentration to be within a specified range. It is therefore necessary to quantify the cleaned PCR product. This was done by running a known volume of the cleaned template against the Life Tech Low Mass ladder, each band is of a specified length and concentration, and comparison of the intensity of the band of DNA to the control enables quantification of the template.

2.4 Cloning

There were some PCR fragments of the viral samples that, for unknown reasons, simply refused to sequence; this led to the need for cloning. The cloned fragments are outlined in Table 3.1. The protocol used was as outlined in the Promega pGEM®-T and pGEM®-T Easy vector systems technical manual.

2.4.1 Ligation

The cleaned and quantified PCR fragments were ligated using the Promega pGEM®-T Easy vector system, in Life Technologies MAX Efficiency DH5α™ Competent Cells. A 2:1 insert vector molar ratio in 10 μl ligations was set up containing 1X T4 buffer, 50 *ng* of Promega pGEM®-T vector, 2 units of t4 ligase, an appropriate amount of PCR product to give the 2:1 molar ratio required, and H_20 to give 10 μl total volume. To allow for complete ligation, the reaction was stored at 4°C overnight.

2.4.2 Transformation

Thirty μl of the MAX Efficiency DH5α™ Competent Cells were used in the transformations. The entire ligation was added to the cells, and maintained on ice for 30 minutes. The cells were heat shocked at 37°C for 40 seconds and then placed back on ice for another 5 minutes. Two hundred micro litres of LB broth was then added, and the cells were shaken at 37°C for one hour. This cell suspension was then plated onto LB -IPTG/X-Gal/Amp plates. The dried plates were inverted and incubated at 37°C overnight. *Lac* operon blue/white colony selection was used to identify transformants.

2.4.3 Recombinant DNA extraction

Possible positives were picked and incubated overnight in 5 *ml* LB broth with 100 *µg/ml* ampicillin on a shaker. These cells were centrifuged at 13 000 rpm for 1-2 minutes then re-suspended in 350 *µl* of STET buffer. To this, 25 *µl* of lysozyme solution was added and the cells were heat shocked in boiling water for 40 seconds. The lysed cells were then spun for 10 minutes at 10 000 rpm, and the debris was aseptically plucked out of the supernatant. The DNA was then extracted using a standardised phenol/chloroform procedure of 0.5 volumes of phenol and 0.5 volumes of chloroform, which is spun for approximately 5 minutes at 13 000 rpm to separate the two phases. The aqueous phase was then washed in 1 volume of chloroform. The DNA was precipitated for one hour at room temperature in a standard salt ethanol solution, then spun for 10 minutes at 13 000 rpm to complete the precipitation. The DNA pellet was washed in 2.5 volumes of 70% ethanol, air dried and re suspended in 25-50 *µl* of H_2O, depending on the size of pellet.

The presence of a PCR insert in the extracted Plasmid DNA was assayed by enzymatic digestion of a sample of the DNA. Ten units of digestion enzyme *Eco*R1, 1X digestion buffer (5mM NaCl, 100mM Tris HCL pH 7.5, 10mM MgCl, 0.025% Triton X--100), 5 *µl* of DNA from the extracted DNA sample and made up to a total volume of 25 *µl* with ddH_20 and incubated at 37°C for one hour. In the last two minutes of digestion 2 *µl* of RNAse solution was added to break down any RNA which interferes with gel staining. Thirty to fifty percent of the digested product was then quantified on 1% agarose gels, to check both for the presence of insert and to quantify the DNA extract.

2.5 Sequencing

Single stranded automated sequencing with ABI Ampli*Taq*® polymerase dye terminator, or dye primer chemistries, were the sequencing methods of choice.

2.5.1 HBV samples

Our analysis of the HBV samples required sequencing only the dominant strain within the individual, consequentially Dye-terminator chemistry was used. This chemistry is more cost effective and less time-consuming to use than the alternative Dye-primer chemistry. A combination of ABI Prism™- Classic Terminator™, dRhodamine, and BigDye™ chemistries were used depending on availability, all of which used the same protocols. 50 *ng* of DNA template was added to 8 *µl* of sequencing reagent and 3.3 *p* mol of primer in a total volume of 20 *µl* made up with ddH_20.

Optimal sequencing reaction required the thermal cycle to undergo specific temperature graduations, thus the protocol used was as follows:

Step 1: +1 °C/second to 96°C
Step 2: 96°C for 10 seconds
Step 3: -1 °C/second to 50°C
Step 4: 50°C for 5 seconds
Step 5: +1 °C/second to 60 °C
Step 6: 60°C for 4 minutes
Step 7: repeat from Step 2 to Step 6 for 24 cycles
Step 8: rapid ramp down to 4°C and hold

The samples were then precipitated and dried using a standardised ethanol precipitation. Precipitation was for 45 minutes at 4°C, followed by a 4°C spin at 14 000 rpm for 30 minutes. The supernatant was then decanted, and the pellet was washed in 0.75 ml of 70% ethanol. After a brief vortex the precipitate is re-spun at 4°C spin at 14 rpm for 15 minutes. Big Dye terminator chemistry however, only required 5 minutes precipitation time and 15 minute spins. The precipitated sample was then dried before scanning on the Massey University ABI 377 genscan.

2.5.2 Trobriand samples

HLA gene analysis techniques require the ability to determine heterozygotes, where the nucleotide sequence has two different bases at one point. Heterozygotes are identified by the presence of two separate peaks at the same position in the electropherogram. Theoretically the two peaks of the same height and half that of the homozygote base peaks.

Of the ABI sequencing chemistries available at the time, only the Dye primer chemistry gave even electropherogram peaks that were necessary to allow heterozygote determination, consequentially it was this chemistry that was employed to determine the HLA sequence. ABI Dye Primer chemistry reactions require labelled primers, and it is possible to have them custom made, but specific primer kits are available which are less expensive. The HLA-DPA1 PCR primers have an M13 tail added to their 5' end, so M13 reverse and -21M13 forward primer cycle sequencing kits could be used. The protocol used was according to the ABI PRISM™ Dye primer cycle sequencing ready reaction kit protocols. Four tubes per DNA sample containing the specified volumes and concentrations were set up as outlined in Table 2.3. The amount of DNA template used was such that a total pool of individual primer reactions gave 40 *ng*. Because a hot top was used in all amplifications, no oil was required to stop evaporation of the PCR components during the reaction.

Table 2.3 ABI dye primer sequencing protocol
There are four solutions in this protocol each with a different labelled base. **A**, **C**, **G** and **T** therefore represent the different labelled nucleotide solutions.

a. M13 reverse classic ready reactions protocol

Reagents	A μl	C μl	G μl	T μl
Ready reaction premix	4	4	8	8
DNA template at concentration of 10 ng/μl	1	1	2	2
total volume	5	5	10	10

b. -21M13 forward BIG dye ready reactions protocol

Reagents	A μl	C μl	G μl	T μl
ready reaction premix	4	4	4	4
DNA template at concentration of 10 ng/μl	1	1	1	1
total volume	5	5	5	5

Optimal sequencing reaction requires the thermal cycle to undergo specific temperature graduations, thus the protocol used was as follows:

Step 1: +1°C/sec to 96°C
Step 2: 96°C for 10 seconds
Step 3: -1°C / sec to 55°C
Step 4: 55°C for 5 seconds
Step 5: +1°/sec to 70°C
Step 6: 70°C for 60 seconds
Step 7: repeat from Step 2 to Step 6 for 15 cycles
Step 8: +1°C/sec to 96°C
Step 9: 96°C for 10 seconds
Step 10: -1°/sec to 70°C
Step 11: 70°C for 60 seconds
Step 12: repeat from Step 8 to Step 11 for 15 cycles
Step 13: rapid rampdown to 4°C and hold

Each sample was then pooled and the DNA sequence was precipitated using the standard method (see above).

2.6 Sequence analysis

The dried samples were then submitted to MuSeq, Massey University's sequencing facility, Palmerston North, for analysis on their 377 ABI Applied Biosystems automated DNA sequencer.

2.6.1 HBV samples

The resulting electropherograms of the sequence data were processed through ABI Sequencing analysis software™ using the Basecall ABI 100 and dRhodamine Terminator or BigDye™ Terminator mobility files (depending on the chemistry used). These sequences were then checked with the MT Navigator PPC software which enabled visualisation of the peaks associated with the base calls. Using an HBV consensus sequence, the entire genomes of the individual strains were also aligned and rechecked in the MT Navigator PPC software.

The genomic HBV sequences of both the HBVs sequenced for this thesis and the seventy obtained from Genbank were aligned in the automatic Clustal X program. This alignment was then checked and completed manually in the program Se-Al (see Table 3.3 for the HBV Pacific samples genomic alignment). The combined alignment of the 16 Pacific samples and the HBV database sequences were then exported for analysis in PAUP * (Swofford *et al.* 1993).

2.6.2 Trobriand samples

The resulting electropherogram of the sequence data was processed through ABI Sequencing analysis software™, using Basecall, ABI 100, and M13 forward or reverse mobility files. Sequence files were then analysed with Match Tools™ (ABI) software by comparing them to a custom made library for the HLA DPA1 gene (created and provided by Zlatibor Velickovic). This software assigns an allele or possible allele combination for each sample (Velickovic *et al.* 1998).

Using MT Navigator software, the test sequences were aligned with the DPA1 exon 2 consensus sequence, and polymorphic positions were re-examined on the electropherogram. Any ambiguities were resolved and sequences were submitted to Match Tools™ (ABI) analysis for final allele typing (Figures 2.2 and 2.3)

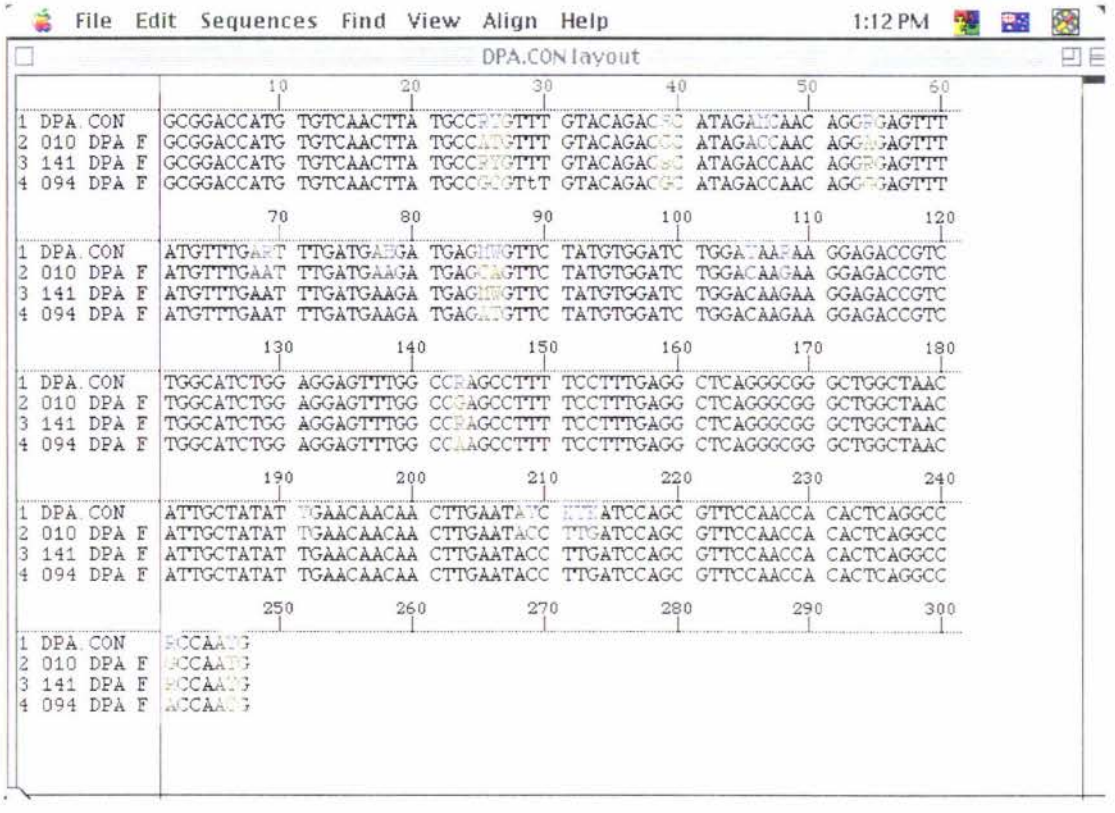


Figure 2.1 Example of MT Navigator PPC alignment of Trobriand HLA-DPA1 alleles showing consensus HLA-DPA1 exon two, and the different allele combinations found in this study. The coding of the degenerate bases matches that in table 2.1a.

Highlighted bases are potential heterogenetic sites.
DPA Con. is the HLA-DPA 1 consensus sequence
Sample 010 is homogenetic for DPA1*02022
Sample 141 is heterogenetic for DPA1*0103/DPA1*02022
Sample 094 is homogenetic for DPA1*0103

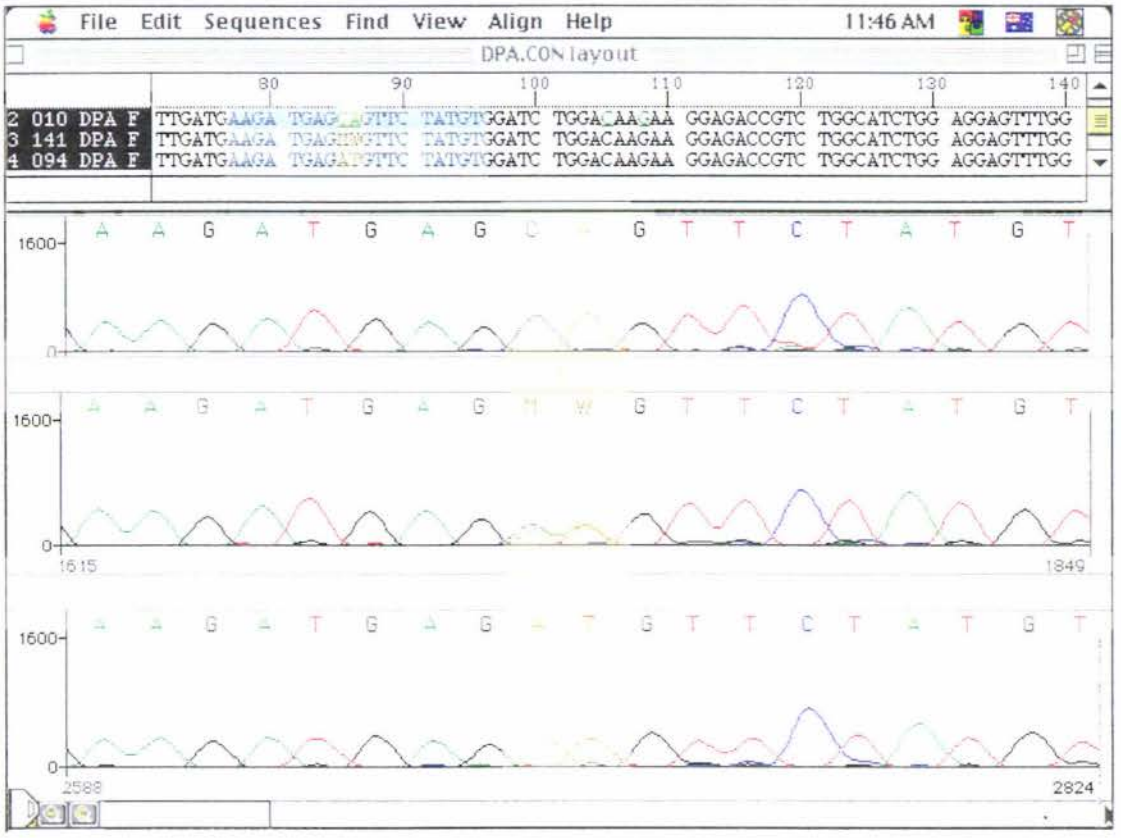


Figure 2.2 Examples of MT Navigator PPC alignment of Trobriand HLA-DPA1 alleles showing the different allele combinations found in this study. Here the corresponding electropherograms are showing. The blue shaded area at the top represents the bases displayed, the highlighted box indicates the heterogenetic sites at position 85 and 86.

at nucleotide positions 85 and 86 :

Sample 010 has bases	C	A
Sample 141 has bases	C & A (M)	A & T (W)
Sample 094 has bases	A	T

Sample: 010 DPA F Library: DPA1.L10 for

Allele Report: Exact match to: DPA1*02022. See Warnings Below.
Files : 010 DPA F

Warning #5: Only one sequence orientation is present.

Warnings for file: 010 DPA F.
#11: The peak spacing of -9.00 falls outside of the normal range of 9.0 to 14.0.

Polymorphic Position Report

```
         1 111222222 | nucleotide
2234567880 049011144 |   number
5696498565 831912316 |

ATCCAAACAC GGTCTTGGT <>   010 DPA F consensus
.......... .........  >   010 DPA F

.......... .........  DPA1*02022
.........T .........  DPA1*02021
.........Y .........  DPA1*02021/02022
```

Sample: 141 DPA F Library: DPA1.L10 for
Allele Report: Exact match to: DPA1*0103/02022. See Warnings Below.

Files : 141 DPA F

Warning #5: Only one sequence orientation is present.

Warnings for file: 141 DPA F.
#11: The peak spacing of 7.00 falls outside of the normal range of 9.0 to 14.0.

Polymorphic Position Report

```
         1 111222222 | nucleotide
2234567880 049011144 |   number
5696498565 831912316 |

RYsCRAAMWC GRTCTTGRY <>   141 DPA F consensus
.......... .........  >   141 DPA F

.......... .........  DPA1*0103/02022
.........Y .........  DPA1*0103/02021
......M... .........  DPA1*0104/02022
......M..Y .........  DPA1*0104/02021
.......... .......GT  DPA1*0105/02022
.........Y .......GT  DPA1*0105/02021
```

```
Sample: DPA 094 F            Library: DPA1.L10 for
Allele Report: Exact match to: DPA1*0103.  See Warnings Below.

Files : 094 DPA F  copy
```

```
Warning #5: Only one sequence orientation is present.
```

```
Polymorphic Position Report

         1 111222222 | nucleotide
2234567880 049011144 |   number
5696498565 831912316 |

GCGCGAAATC GATCTTGAC <>   DPA 094 F consensus
.......... .........  >   094 DPA F  copy

.......... .........  DPA1*0103
......M... .........  DPA1*0103/0104
......C... .........  DPA1*0104
.......... .......RY  DPA1*0103/0105
.......... .......GT  DPA1*0105
......M... .......RY  DPA1*0104/0105
```

Figure 2.3 ABI Match Tools final analysis out put file for the Trobriand HLA-DPA1 alleles. Examples are of samples 010, 141, and 094. The program identifies the allele as well as other closely related allele combinations, for example, sample 010 is an exact match to DPA1*02022, with the closest alternatives being DPA1*02021 or DPA1*02021/02022

2.7 Phylogenetic analysis

Phylogenetic reconstruction (evolutionary tree building) is the process of inferring the "best estimate" of the evolutionary relationship among a group of taxa (in this case sequences). Irrespective of the aim, whether it be recovering the tree of relationships, or estimating rates of change and/or times of divergence, the tree reconstruction method can be characterised by the following components:

- The data, its selection and preparation.
- Transformations, such as any correction for multiple changes.
- Tree selection criterion (optimization).
- The search strategy (exact or heuristic).
- Testing and evaluation of the inferred trees (Penny *et al.* 1996; Swofford *et al.* 1996).

There is a multitude of methods available for use in obtaining the best estimate of phylogenetic relationships. Outlined below is a brief description of the methods used in this study.

2.7.1 The data

Phylogenetic analysis of sequence data requires positional homology between the sequences — that is, it should be possible to trace the ancestry of nucleotides observed at a given site in the sequences, to a single position that occurred in a common ancestor of those taxa. In order to meet this requirement, gaps (corresponding to insertions and/or deletions) are inserted into the sequences, creating an alignment. The alignment algorithms are designed to maximise sequence similarity while minimising the number of insertions and deletions (indels). Consequently, base substitutions (where two nucleotides do not match in the aligned sequence) are assumed to be more frequent and are penalised less severely by the alignment algorithm than gaps (Swofford *et al.* 1996).

Most automated multiple alignment programs first make pairwise alignments and then combine the sequences together by inserting additional gaps as required. As a consequence the order of the sequences in the data set can effect the final alignment. Clustal X (Thompson *et al.* 1994) attempts to overcomes this potential bias by obtaining the order of pairwise alignment from clusters in an initial tree produced from a matrix of distances across pairwise alignments. This creates an alignment that is not order dependant, but is dependant on this initial tree (Thompson*et al.* 1994; Swofford *et al.* 1996).

Visual inspection and correction is still needed in many cases to ensure a sensible alignment has been obtained. Therefore, as a matter of course throughout this project all sequences were aligned in Clustal X then manually checked in Se-Al. All columns with gaps were removed for phylogenetic analysis, as they cannot be accurately accounted for in most analyses (Swofford *et al.* 1996).

2.7.2 Tree Reconstruction

The perfect tree-building algorithm is one that is :

- *Efficient* : Is computationally fast (for instance, if computation time increases less than exponentially, as more taxa are added and/or longer sequences are used).
- *Consistent* : Refers to the ability of the method to 'converge' upon the tree used to generate the data, when longer and longer sequences are used.
- *Powerful* : The shorter the sequence required to converge to the correct tree, the more powerful the method.
- *Robust* : Under one set of prescribed assumptions a method may be consistent, but may lose consistency when the true conditions deviate from those assumptions. A method is robust when it continues to converge to the correct tree even when the underlying model deviates from the assumptions.
- *Falsifiable* : Refers to the method being able to reject the model, it is satisfied when the method can do so (Penny *et al.* 1999).

As yet this golden egg has not hatched! Faster methods tend to be inconsistent and weak, and the more powerful methods are more computationally expensive and slow. While longer sequence data can aid in obtaining a more accurate tree, the number of sequences (taxa) used in this study limits methods of search strategy analysis. Exact searches using branch and bound use a prohibitive amount of computing time for more than about 20 sequences (Hendy and Penny 1982; Penny *et al.* 1992; Swofford *et al.* 1996). In addition many methods, while mathematically tractable, have biological flaws. For example, most models assume that the bases evolve identically and independently throughout the molecules, but this assumption causes problems when considering functional proteins.

Therefore, in an attempt to obtain the "best estimate" of the phylogenetic relationship for the HBV sequences, a combination of optimality criteria, evolutionary models and statistical evaluations was used.

2.7.3 Models: transformations

Reconstructing evolutionary trees depends on a model of evolution, which consists of three parts:

- A tree (or, more generally, a graph, which allows recombination, lateral transfer and or hybridisation).
- A mechanism of evolution (stochastic changes, identically independent distributed (iid), frequencies of classes of mutation).
- Initial conditions, rates of evolution.

For example, the Jukes and Cantor (1969) model assumes equal base frequencies with a single substitution type, whereas the Hasegawa-Kishino-Yano (HKY85) model (Hasegawa *et al.* 1985) allows an unequal equilibrium of base frequencies, with different rates of transitions and transversions. Rate heterogeneity between sites (e.g. a gamma distribution across sites) can also be incorporated into the analysis. Due to the composition of the HBV genome, HKY85 was always used in maximum likelihood and neighbour-joining analysis (Swofford *et al.* 1996).

Covarion model of evolution

"Covarion" is a contraction of **co**ncomitably **vari**able cod**on**s (Penny *et al.* 1999). The basis of the model is that while some sites are unable to change ('fixed') due to their function within a molecule, others switch between being free to evolve, and being fixed. The changes from 'fixed', to 'free to vary', will depend on slight changes in the 3-D structure during evolution.

These covarion patterns can mislead evolutionary tree building when distantly related sequences show more similar distributions of invariable sites than closely related taxa. While the model is biologically acceptable it is mathematically complex. At present there are no models of covarion evolution incorporated in available phylogenetic software packages. There is however an inequality test available that can reject or accept an underling difference in covarion structure (Lockhart *et al.* 1998).

2.7.4 Tree Selection criteria

2.7.4.1 Optimality Criteria

Methods which use an optimality criteria evaluate this criteria on as many trees as possible and attempt to find the tree(s) which optimise this criterion. An optimality criteria which is global evaluates all combinations of taxa, then selects the optimal tree(s) defined by the objective function for that criteria. They are computationally slow, in fact, prohibitive with exact or even branch and bound searches of more than a few taxa (about 20 taxa for parsimony and about 10 taxa for maximum likelihood) (Hendy and Penny 1982; Penny *et al.* 1992; Swofford *et al.* 1996).

Parsimony

The maximum parsimony criterion minimises the number of changes (mutations), be they nucleotide or amino acid, required to explain the observed data on a given tree(s). As currently implemented, parsimony does not usually use data corrected for multiple changes. Without corrections for multiple changes parsimony does not guarantee a correct result but gives an explanation avoiding "ad hoc" hypotheses whenever possible. Moreover, under well known conditions, such as rate heterogeneity for even 4 taxa, uncorrected parsimony has been shown to be inconsistent and is susceptible to long branch attraction (where taxa are placed together not due to inheritance but rather because of parallel changes) (Swofford *et al.* 1996). This needs to be kept in mind in the proceeding analysis when considering the mutation ability of the virus, and the long branch of the HBV-**F** genotype.

Maximum likelihood.

Maximum likelihood (Felsenstein 1981) evaluates the probability of observing the data, given a proposed model of evolution. It does so using an iid model, where the log-likelihood of all possible scenarios are summed at each site, and then summed again, over all sites to give the log-likelihood of the tree. An hypothesis is chosen that maximises the probability of observing the data (Felsenstein 1981; Swofford *et al.* 1996). In addition, maximum likelihood treats changes as more likely along longer branches than short ones, so estimation of branch lengths is an important component of the method. This difference explains, in part, the greater consistency of maximum likelihood for sequence data compared to parsimony, which presently does not take into account information on branch lengths when evaluating a tree.

Another advantage of maximum likelihood is that it tends to be robust to many violations of the assumptions used in its methods. How the probabilities of the various changes are calculated depends on assumptions of the process of nucleotide substitution, i.e., the model of evolution (see 2.7.3) (Swofford *et al.* 1996).

2.7.4.2 Constructive

Constructive methods (usually based on distances) use an algorithm to build up the complete tree by joining pairs of taxa. They are a quantitative composition of the data, suboptimal trees are not evaluated and there is no indication of other equi-probable tree topologies. Corrected distances account for superimposed changes by estimating the number of unseen events, using the same models/mechanisms employed by maximum likelihood (see 2.7.3). These corrected distances are estimates of the true evolutionary distance between taxa since divergence. Information is lost though in the conversion of character state to pair-wise distance (that is percent similarity or dissimilarity). Distance based methods are not as generally consistent as maximum likelihood methods, but they have increased computational speed, and consequently they are of use in large data sets were maximum likelihood is computationally prohibitive.

Split decomposition

Split decomposition is a method for graphically representing patterns in distance data, and does not assume that the data will conform to a tree. While detecting well-supported 'real' groupings when present, it also identifies conflicting information that may also have strong support in the data. These conflicts may be due to inadequate correction for superimposed changes in the distance transformation, or conversion (parallel changes) due to natural selection. This method allows simultaneous examination of phylogenetic signals and competing hypotheses (Bandelt and Dress 1992).

2.7.5 Search strategies

2.7.5.1 Global

Exact: Complete search and Branch and Bound

Exact algorithms effectively consider all trees, either by an exhaustive search, or by branch and bound search strategies. Branch and bound eliminates parts of the search space that contain only sub-optimal solutions, so although it is not exhaustive it reaches an optimal answer with less computation. Branch and bound currently is prohibitive with more than about 20 taxa for parsimony or closest tree (Hendy and Penny 1982; Swofford *et al.* 1996).

Heuristic

Heuristic approaches are applied when the data set is too large for exact methods, in which case optimal trees are not guaranteed. Commonly a method starts with an initial tree, which is then rearranged successively with the intent to improve the tree. When no more improvement can be obtained under the defined criteria, rearrangement stops. However, whether the tree is globally optimal or just locally optimal cannot generally be determined (Swofford *et al.* 1996).

2.7.5.2 Local

Algorithmic based method

Purely algorithmic search strategies have a fundamental difference from the methods employed in criterion-based methods. In these cases the algorithm is of fundamental importance, defining the tree selection criterion, which contrasts with criterion-based methods, that use the algorithms as a tool in evaluating the objective function. Algorithmic-based methods define a single tree without reference to any others of equal or nearly as much support (Swofford *et al.* 1996).

Neighbour-joining

Neighbour-joining (Saitou and Nei, 1987; Studier 1988) is a distance-based method which does not assume that all lineages have diverged by equal amounts from each other, that is, it does not assume a molecular clock. It does however assume the data comes close to fitting an additive tree, and consequently correction for superimposed substitutions is important. The raw data is converted to a distance matrix, which is then modified by the average divergence between pairs of taxa. The tree is constructed in an additive manner by first linking the strongest pair creating a cluster. The ancestral node of this cluster is then included in the data set, and the process repeated (Swofford *et al.* 1996).

2.7.6 Tree testing and evaluation

Bootstrapping

Bootstrapping (Efron 1982) is a general statistical technique for estimating the reliability of a parameter when only a single dataset is available. For phylogeny it resamples by drawing sites, with replacement, from the original dataset thus creating multiple new datasets of the same size. A tree is formed from each new dataset, the individual properties are recorded, and used to create a consensus tree or a list of most common internal branches. From this consensus tree and list, estimates of the confidence limits of internal branches can be made. Bootstrap values give the measure of support there is in the data for a given criterion. Bootstrapping does not confirm that the given tree is correct, but rather high bootstrap values mean the patterns are present in the data and adding more similar data should not markedly alter the best tree — so longer sequences will not change the choice of optimal tree for the given criterion (Efron 1982; Swofford *et al.* 1996).

Nearest Neighbour bootstrap

Whereas bootstrapping gives a measure of the tree stability, the nearest neighbour bootstrap (NNB) method gives a measure of the local branch stability of the optimal tree (Matisso-Smith *et al.* 1998). NNB indicates the degree of support for a tree if minor local variations in taxon placement are permitted. It allows two rearrangements around every internal branch of a tree that are not in the consensus tree, it sums the bootstrap values to give a total percent support for the consensus tree and the combined one step rearrangements (Matisso-Smith *et al.* 1998).

Covarion Inequality test

The Covarion Inequality test (Lockhart *et al.* 1998) compares the proportion of constant and variable sites between two groups (eg. genotypes) (see Table 2.4). It takes into account the following: The number of invariant sites across both groups of taxa (N1). The number of sites that are invariant in both groups of taxa but are different between the groups (N2). The number of variable sites in the first group that are invariable in the second (N3). The number of variable sites in the second group that are invariable in the in the first (N4), and finally the number of sites that are variable in both groups (N5) .

Table 2.4 Covarion Inequality test character descriptions

	N 1	N 2	N 3	N 4	N 5
Group 1	X	X	X	X	Y
	X	X	Y	X	X
	X	X	X	X	Y
	X	X	Y	X	X
Group 2	X	Y	X	Y	Y
	X	Y	X	X	X
	X	Y	X	Y	Y
	X	Y	X	X	X

This test compares the observed N5, with the expected N5. This is calculated from the product of the probability of a given site being variable in group one (N3+N5), and that of a given site being variable in a second group (N4+N5), divided the total number of sites (N3+N5)(N4+N5)/N (N being the total number of sites). In the covarion models studied so far N5 will be less than expected, since a site which is varied in one group is less likely to be varied in another group, whereas under a rates across sites model (a model which allows some sites to be constant and other sites to have different rates) a site which is varies in the first group is more likely to vary in the second (Lockhart *et al.* 1998).

This test uses SplitsTree to give the required (N, N1-5) which are calculated by using the 'group-distance function'. These values are then incorporated into an Excel program provided by P. Lockhart, IMBS, Massey University and M. Steel, University of Canterbury. This program calculates the mean, standard deviation and Z values. A 'Z' value of less than -2 is considered significant and indicative of a change in covarion structure of the molecule during evolution

2.8 Software packages and Data bases

Table 2.5 Software packages and Data bases used with in this study.

Software	Function	Source	Reference
Clustal X	Alignment	http://www.csc.fi/molbio/progs/ clustalw/clustalw.html #DIST	Thompson *et al.* (1994)
Genbank	Sequence Database	http://www2.ncbi.nlm.nih.gov/cgi-bin/genbank	n/a
Match Tools ™	Compares sequences to custom library.	Perkin Elmer Applied Biosystems http://www.perkin-elmer.com/ab	n/a
Mega	Phylogenetic analysis and data presentation	Imeg@psuvm.psu.edu	Kumar *et al.* (1993)
MT Navigator PPC	Electropherogram analysis. Alignment	Perkin Elmer Applied Biosystems http://www.perkin-elmer.com/ab	n/a
NewBoots	Determines local tree stability	D. Penny, Institute of Molecular BioSciences, Massey University, New Zealand.	Matisoo-Smith *et al.* (1998) D.Penny @massey.ac.nz
Oligo®4.03	Primer design	National Biosciences Inc. 3650 Annapolis Lane. Plymouth, MN 55447. USA.	n/a
PAUP* 4.0.0d64 (Phylogenetic analysis using Parsimony)	Phylogenetic analysis	http://www.sinauer.com	Swafford (1993)
PHYLIP	Phylogenetic analysis	http:evolution.genetics.washington.edu/phylip	Felsenstein (1993)
Se-Al	Manual alignment	http://evolve.zps.ox.ac.uk/ Se-Al/Se-Al.html	Andrew.Rambaut @zoo.ox.ac.uk
Splits Tree	Phylogenetic analysis	Huson@mathematik.unibielefeld.de ftp://ftp.uni-bielefeld.de/pub/math/ splits	Huson (1998)

Chapter Three: **Results**

The data produced, and its analysis is described in this chapter in three sections. It begins with the genomic sequences obtained from the HBV-positive serum samples, their features, and phylogenetic analysis. This is followed by the complete HBV genomes collected from Genbank and the phylogenetic analysis of a selection of these samples combined with our Pacific samples. This section includes the hypothesis testing for variation in covarion structure between genotypes. Finally the HLA-DPA sequences and typing are considered.

3.1 Pacific HBV samples; sequences, features, and phylogenetic analysis

3.1.1 HBV sequencing results

Sixteen HBV positive serum samples were provided from which 14 complete genome sequences and 2 partial sequences were derived. The mutation rate of HBV is high; consequentially it exists within a host in a quasi-species distribution. A quasi-species distribution is a population of viruses that share a common origin but which have distinct genomic sequences as a result of mutation, drift, and the impact of selection. A single individual is therefore infected with many distinct variants that together are a quasi-species (or form a quasi-species distribution) (Smith *et al.* 1997). The nature of PCR enables all these variants to be amplified, however the ABI dye terminator chemistry available at the time only enabled definitive reading of the predominant sequence, so the sequences obtained are these. The 14 complete genomic and 2 partial nucleotide sequences are displayed in Table 3.3. The sequences are aligned using the numbering system of Ono *et al.* (1983) which starts at the *Eco* R1 site. The amino acid sequences for each gene are given in Appendix A.

Routine PCR amplification and sequencing, as described in chapter two, was successful for 10 samples and complete genomes were obtained. However, for six samples there were some troublesome fragments. These gave PCR products, but forward or reverse (or both) sequencing primers would not give a readable sequence, hence they had to be cloned and then sequenced (Table 3.1) The cloned sequences are as given in Table 3.1. A few of the cloned PCR products demonstrated variation in sequence, which is to be expected with viral quasi-species (Alexopoulou *et al. 1997*; Smith *et al.* 1997). The clone sequence that matched the PCR product sequence in the overlapping regions were chosen to complete the genomic sequence, and this resulted in 13 complete genomic sequences.

Table 3.1 Cloned fragments of the Pacific hepatitis B samples.
For each sample cloned the corresponding fragment and primers used to get the fragment are given. For each sample two clones were sequenced in both the forward and reverse direction.

Sample	Fragment position (nucleotide position)	Primer combination	# Clones	Sequence results			
				Forward		Reverse	
				Cl1	Cl 2	Cl1	Cl 2
NZM-1	1655-2304	HB5F-HB5R	2	✓	✓	✓	✓
NZM-4	1307-2304	HB3F-HB4R1	2	x	x	x	x
	1655-2304	HB5F-HB5R	2	x	x	x	x
NZM-6	1307-2304	HB3F-HB4R1	2	x	x	x	x
	1655-2304	HB5F-HB5R	2	x	x	x	x
NZM-5	2446-3065	HB7F-HB7R	2	✓	✓	✓	✓
NZM-8	2667-2667	HB6F-HB6R	2	✓	✓	✓	✓
	2446-0050	HB7F-HB8R	2	✓	✓	✓	✓
NZM-Ma	1655-2304	HB5F-HB5R	2	✓	✓	✓	✓
	2057-2667	HB6F-HB6R	2	✓	✓	✓	✓

✓ = sequencing successful, see results.
X= sequencing unsuccessful
Cl= clone.

The 14th sequence (sample NZM-8) gave interesting results; heterogeneous sequences matching both Genotype **D** and Genotype **C** were routinely obtained. Initially this was considered to be possible contamination, but this was rejected when noncontaminated re-extraction and amplification gave the same results. The PCR amplified fragment that included the preS I region for this sample always gave two separate bands on an agarose gel however stringent the PCR amplification. A specific feature of HBV-**D** genotypes is their 33 bp deletion in the preS I region (bases 2862-2894). This double banding is consistent with the presence and absence of such a deletion. This result, combined with the heterogeneic sequence profile of the sample suggested that the donor was infected with two genotypes. A similar case has been previously described by Bollyky *et al.* (1996). The sequence displayed is one that matches that of the HBV-**C** genotype.

Cloning provided clear sequence for all samples except NZM-4 and NZM-6. The reason why there were problems associated with these two samples was never completely determined. In an attempt to obtain the missing sequence, secondary structure formation was considered where a range of DMSO concentrations and primer annealing temperatures were tested. Also the insertions, deletions, and primer mis-matches were considered. Variations in elongation times, primer annealing temperatures, and primer combinations, were all tried.

However any attempt to amplify any fragment between base positions 1307-2304 met with problems such as no amplified product, multiple banding or, when a fragment of the correct size was purified from a gel cut, unreadable sequence was produced. Cloning was attempted and while positive insert results were obtained, garbage sequence was produced. Table 3.3 shows the sequences produced from the first two primer combinations giving bases 1-974 for sample NZM-4 and 1-910 for sample NZM-6 and the other 14 complete genomic sequences checked, rechecked and checked again !

To enable a clinical evaluation of the sequences background data of the donating individuals was collected. This information is presented in Table 3.2.

Table 3.2 Background information on the hepatitis B serum donors.
The exact age for the last six women is not known accept that the were within child bearing age, and the HBV genotypes which were ascertained in this study is given for completion of the table.

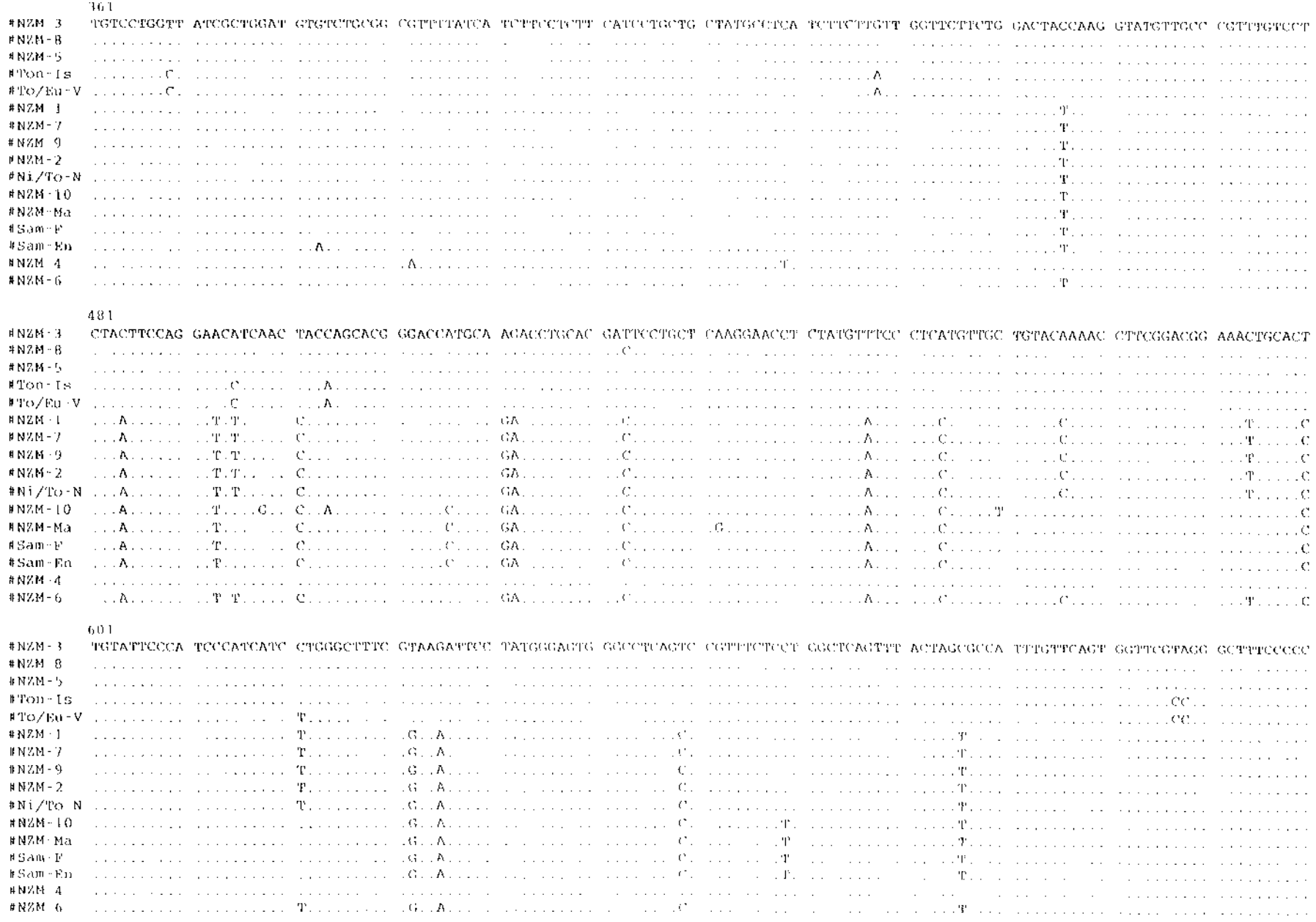
Sample	Age	Sex	Ethnicity	HBeAg (+/-)	ALT	Genotype
NZM-1	21	Male	Maori	+	78	HBV- **D**
NZM-2	19	Male	Maori	+	49	HBV- **D**
NZM-3	22	Male	Pacific Is.	-	63	HBV-**C**
NZM-4	33	Male	Maori	-	36	HBV-**C**
NZM-5	34	Male	Maori	-	27	NZM-**C**
NZM-6	33	Male	Maori	-	34	HBV- **D**
NZM-7	34	Female	Maori	-	198	HBV- **D**
NZM-8	51	Male	Maori	+	70	HBV-**C**/ **D**
NZM-9	34	Male	Maori	-	131	HBV- **D**
NZM-10	39	Male	Maori	-	78	HBV- **D**
NZM-Ma	18-35	Female	Maori	+	16	HBV- **D**
Ni/To-N	18-35	Female	Nuiean/Tongan	+	11	HBV-**D**
Sam-En	18-35	Female	Samoan	+	20	HBV-**D**
Sam-F	18-35	Female	Samoan	+	14	HBV-**D**
To/Eu-V	18-35	Female	Tongan	+	12	HBV-**C**
Ton-Is	18-35	Female	Tongan	+	22	HBV-**C**

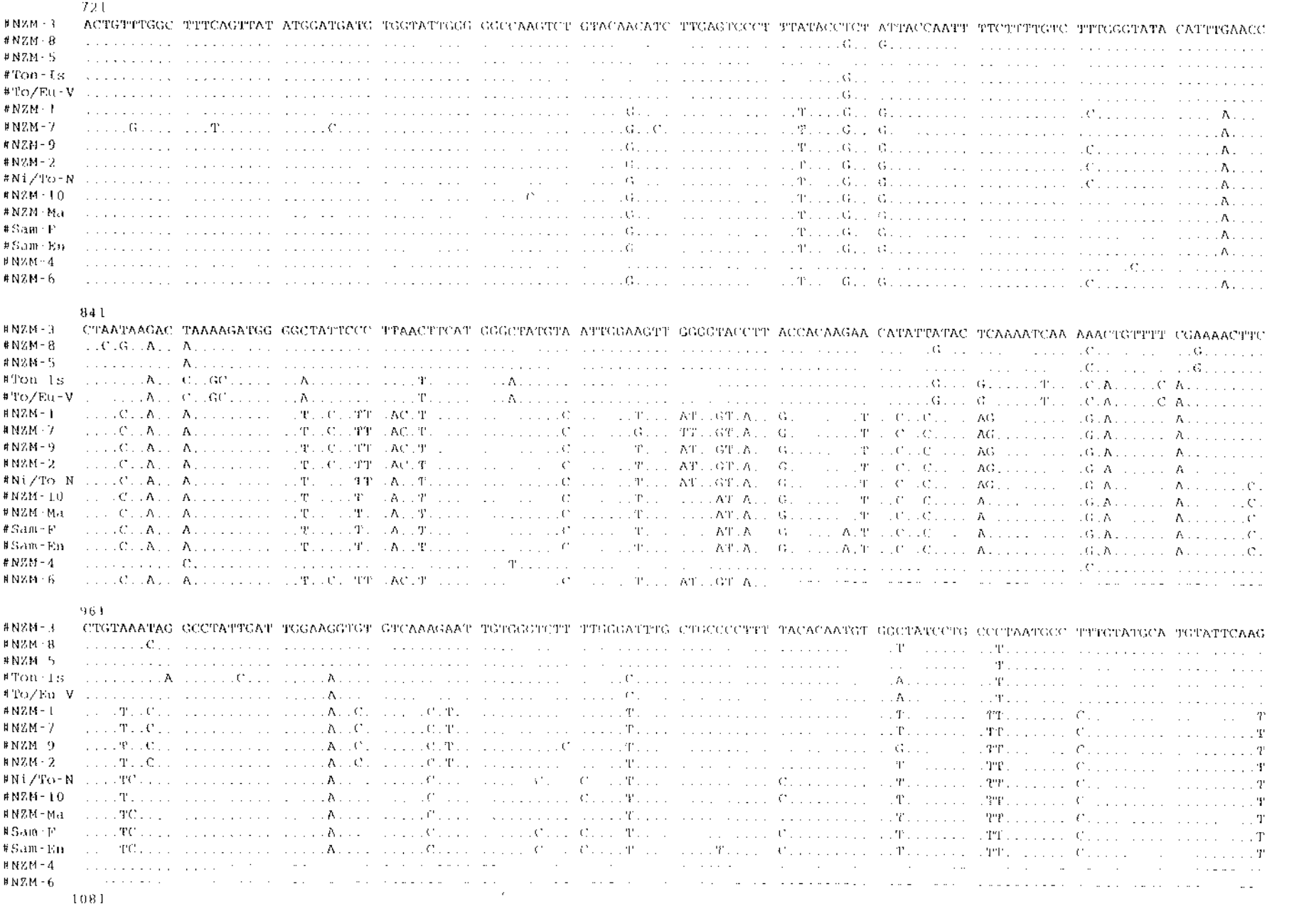
Table 3.3 The genomic sequences of the Pacific hepatitis B virus samples

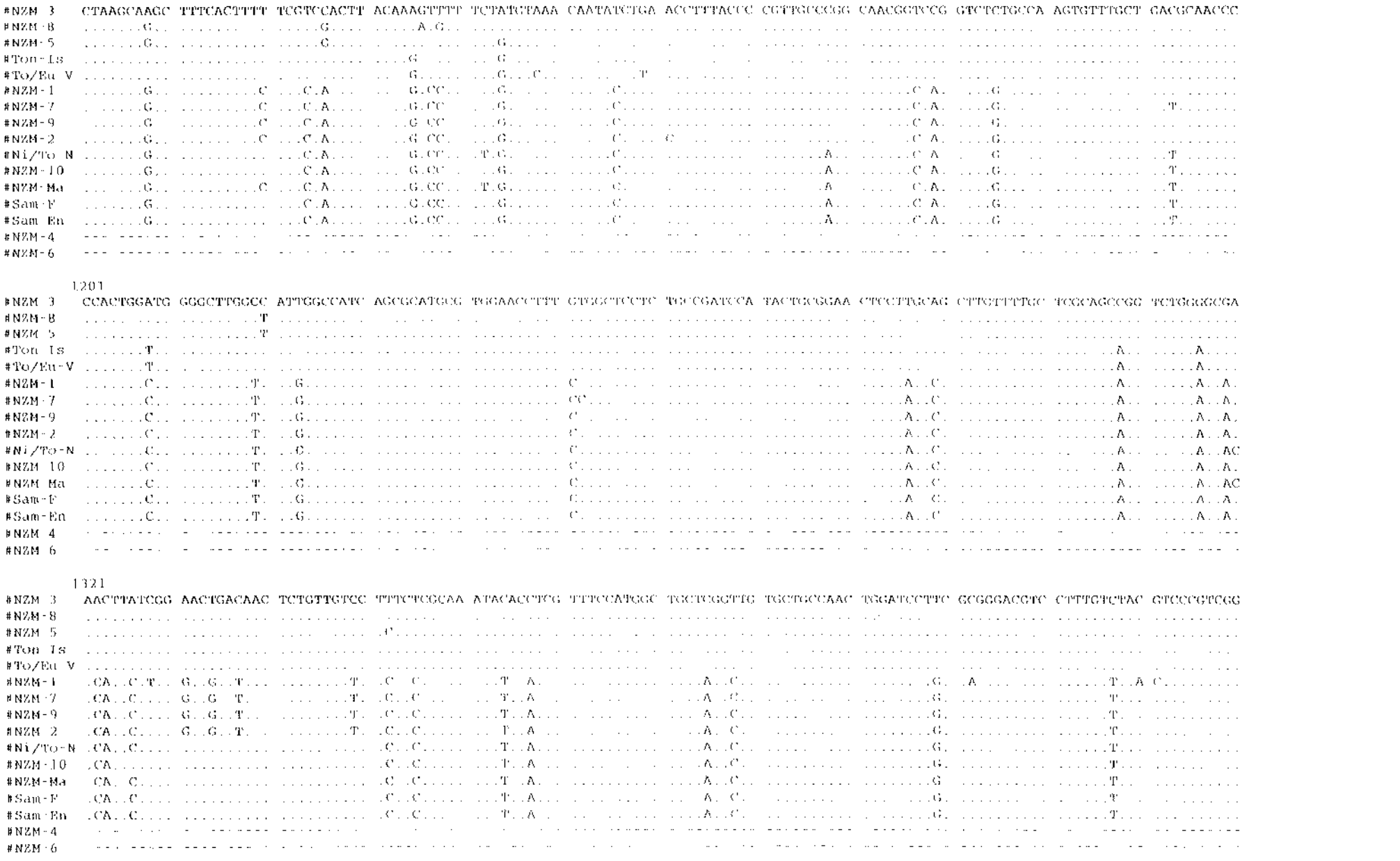
	1											
#NZM-3	AATTCCACAA	CATTCCACCA	AGCTCTGCTA	GATCCCAGAG	TAACGGGCCT	GTATCTTCCT	GCTGGTGGCT	CCAGTTCCGG	AACAGTAAAC	CCTGTTCCGA	CTACTGCCTC	TCCCATATCG
#NZM-8	....	....	....	....	....	....T.....	....	....	....	....	....	....
#NZM-5	.........G	....	....	....	....	....T.....	....	....	....	....	....	....
#Ton-Is	..C.......	....	....	....	.G........	T...TA....	....	....	....	....	....	....
#To/Eu-V	..C.......	....	....	........G.	.G........	T...TA....	....	....	....	....	....	....
#NZM-1	..C.......	.C........	.A.G....A.	....	.G..A.....	....T.C...	....	.......A..	....	....	......T...	..A.......
#NZM-7	..C.......	.C....G...	.A.C....A.	....	.G..AA....	....T.C...	....	.......A..	....	....	......T...	..A.......
#NZM-9	..C.......	.C........	.A.G....A.	....	.G..A.....	....T.C...	....	.......A..	....	....	......T...	..A.......
#NZM-2	..C.......	.C........	.A.G....A.	....	.G..A.....	....T.C...	....	.......A..	....	....	......T...	..A.......
#Ni/To-N	..C.......	.C........	.A.G....A.	....	.G..A.....	....T.C...	....	.......A..	....	....	......T...	..A.......
#NZM-10	..C.......	.C.....T..	.A......A.	....	.G..A.....	....T.C...	....	.......A..	....	....	....	....
#NZM-Ma	..C.......	.C........	.A......A.	....	.G..A.....	....T.C...	....	.......A..	....	....	....	....
#Sam-F	..C.......	.C........	.A......A.	....	.G..A.....	....T.C...	....	.......A..	....	....	....	....
#Sam-En	..C.......	.C........	.A......A.	....	.G..A.....	....T.C...	....	.......A..	....	....	....	....
#NZM-4	....	....	....	....	....	....T.....	....	....	....	....	....	....
#NZM-6	..C.......	.C........	.A.G....A.	....	.G..A.....	....T.C...	....	.......A..	....	....	......T...	..A.......

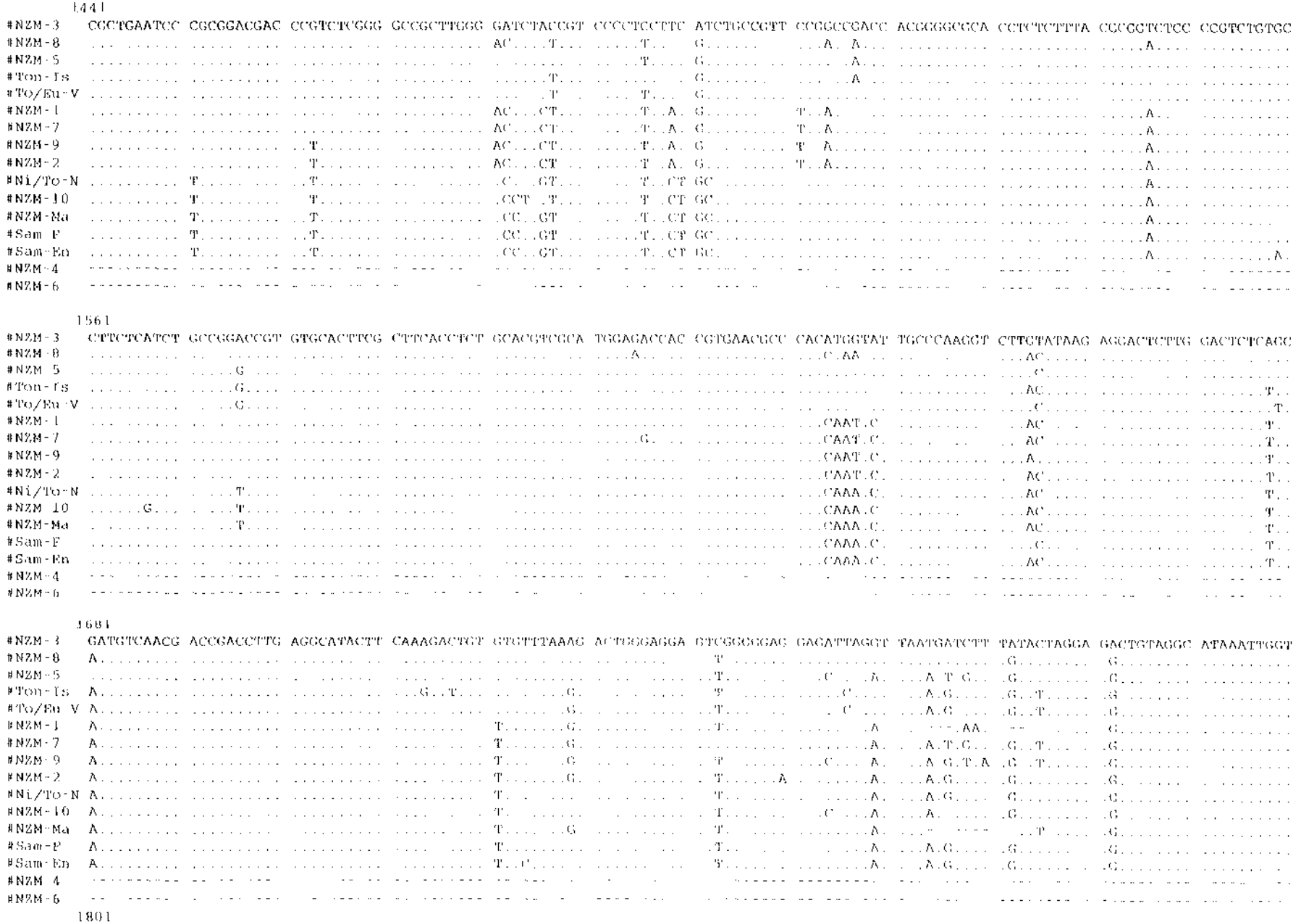
	121											
#NZM-3	TCAATCTTCT	CGAGGACTGG	GGACCCTGCA	CCGAACATGG	AGAACACAAC	ATCAGGATTC	CTAGGACCCC	TGCTCGTGTT	ACAGGCGGTG	TTTTTCTTGT	TGACAAGAAT	CCTCACAATA
#NZM-8	....	....	....	....	....	....	....	....	....	....	....	....
#NZM-5	....C.....	....	....	....	....	....	....	....	....	....	......A...	....
#Ton-Is	....	....	....	....	........G.	....	....	....	....	....	....	....
#To/Eu-V	....	....	....	....	........G.	....	....	....	....	....	....	....
#NZM-1	....	......T...	.........G	.T........	......TC..	....	....	.......A..	........G.	....	....	....
#NZM-7	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#NZM-9	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#NZM-2	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#Ni/To-N	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#NZM-10	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#NZM-Ma	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#Sam-F	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#Sam-En	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....
#NZM-4	....	....	....	....	....	....	....	....	....	....	....	....
#NZM-6	....	......T...	.........G	.T........	......TC..	....	....	....	........G.	....	....	....

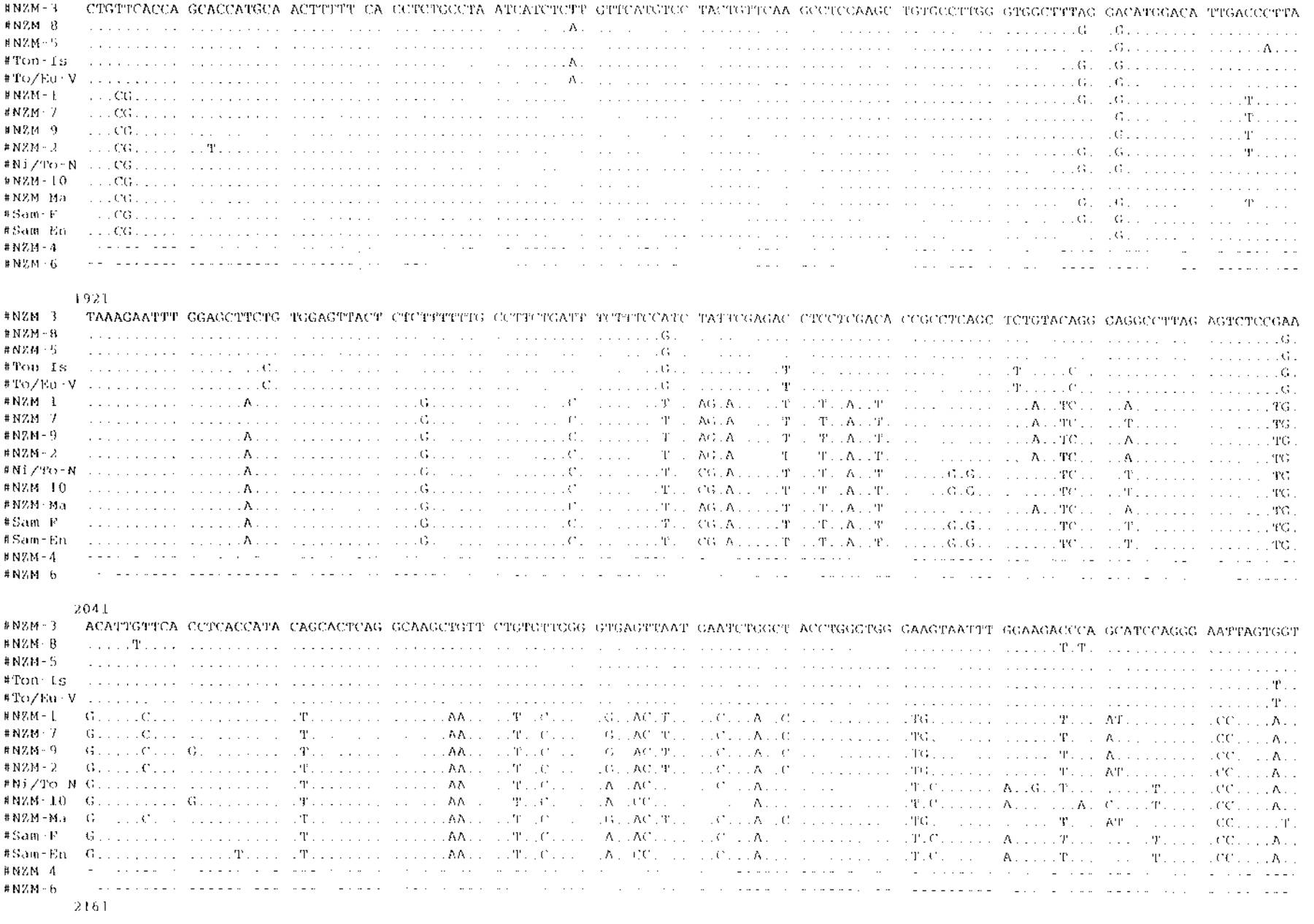
	241											
#NZM-3	CCACAGAGTC	TAGACTCGTG	GTGGACTTCT	CTCAATTTTC	TAGGGGAAGC	ACCAAGGTGT	CCTGGCCAAA	ATTGGCAGTC	CCCAACCTCC	AATCACTCAC	CAACCTCTTG	TCCTCCAATT
#NZM-8	....	....	....	....	....	....	....	....	....	....	....	....
#NZM-5	....	....	....	....	....	....	....	....	....	....	....	....
#Ton-Is	....	....	....	....	....	...CGT....	....	....	....	....	....	........C.
#To/Eu-V	....	....	....	....	....	...CGT....	....	....	....	....	....	........C.
#NZM-1	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#NZM-7	..G.......	....	....	....	......G.A.	TA.CGT....	.G........	....	....	....	.......C..	........C.
#NZM-9	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#NZM-2	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#Ni/To-N	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#NZM-10	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#NZM-Ma	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#Sam-F	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#Sam-En	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.
#NZM-4	....	....	....	....G.....	....	.....A....	....	....	....	....	....	....
#NZM-6	..G.......	....	....	....	......G.A.	TA.CGT....	.T........	....	....	....	.......C..	........C.

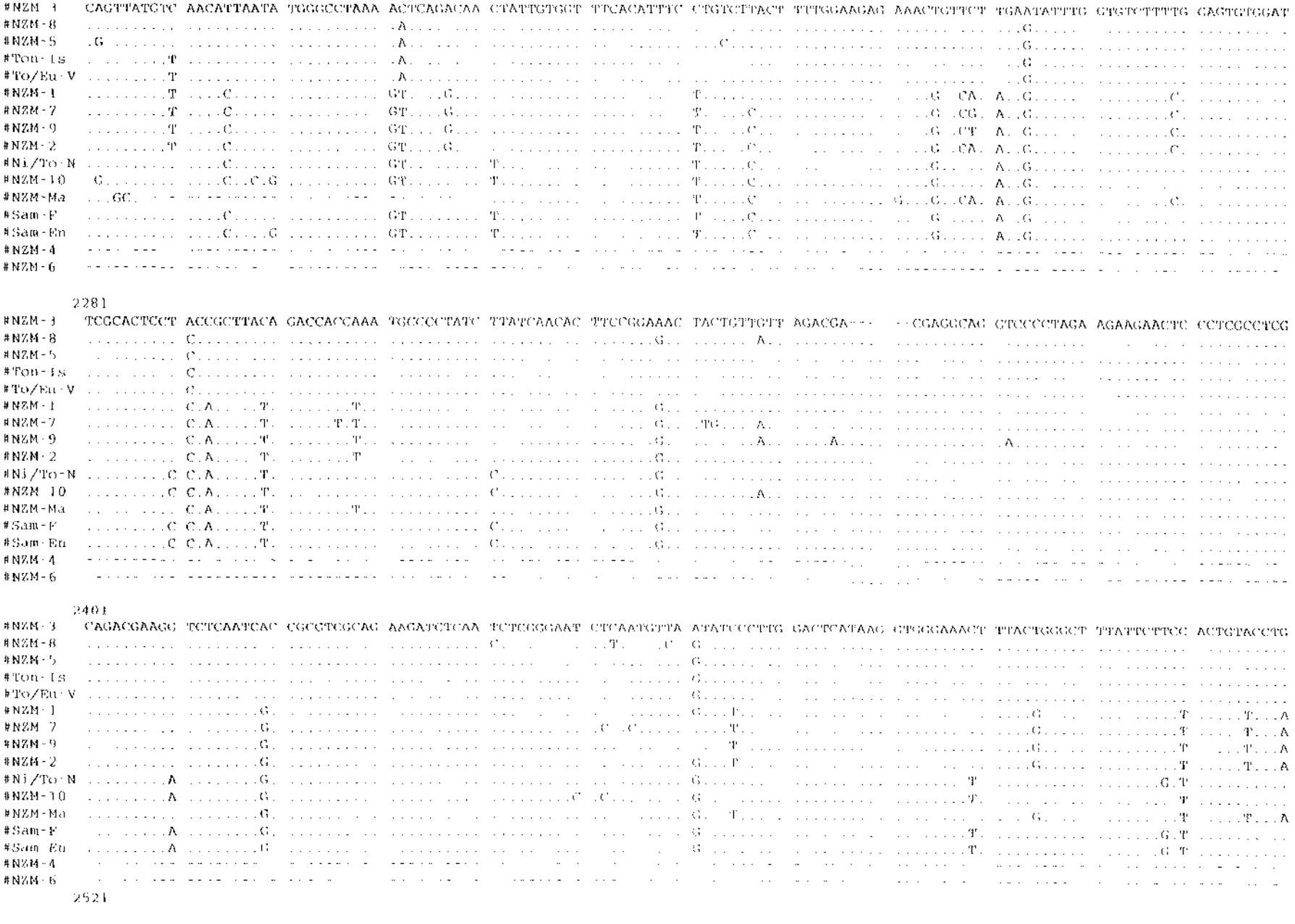

361
#NZM_3 TGTCCTGGTT ATCGCTGGAT GTGTCTGCGG CGTTTTATCA TCTTCCTCTT CATCCTGCTG CTATGCCTCA TCTTCTTGTT GGTTCTTCTG GACTACCAAG GTATGTTGCC CGTTTGTCCT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM_1
#NZM-7
#NZM_9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM_4
#NZM-6
481
#NZM-3 CTACTTCCAG GAACATCAAC TACCAGCACG GGACCATGCA AGACCTGCAC GATTCCTGCT CAAGGAACCT CTATGTTTCC CTCATGTTGC TGTACAAAAC CTTCGGACGG AAACTGCACT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
601
#NZM-3 TGTATTCCCA TCCCATCATC CTCGGCTTTC GTAAGATTCC TATGGGAGTG GCCCTCAGTC CGTTTCTCCT CGCTCAGTTT ACTAGCGCCA TTTGTTCAGT GGTTCGTAGG GCTTTCCCCC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6

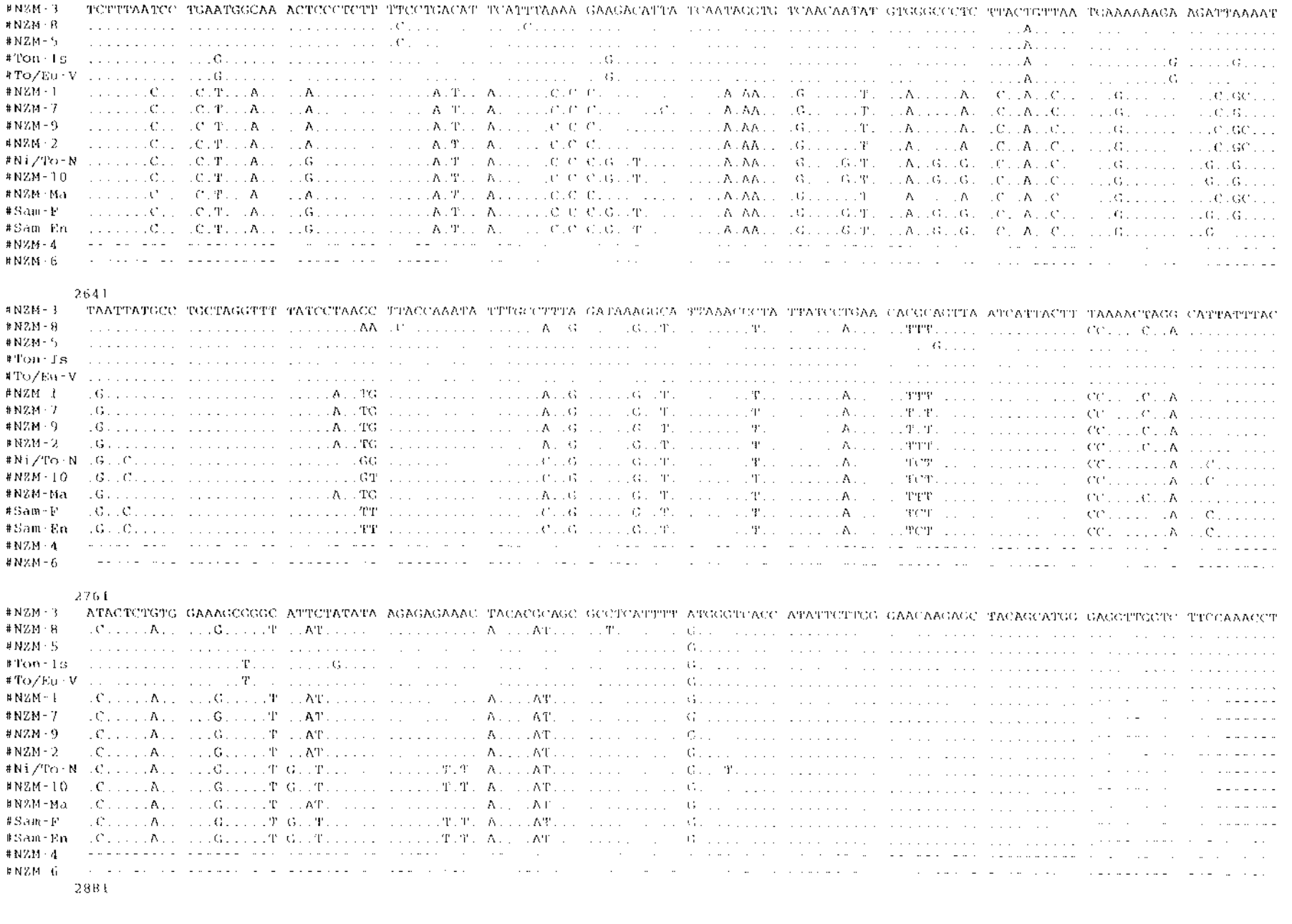

721
#NZM-3 ACTGTTTGGC TTTCAGTTAT ATGGATGATG TGGTATTGGG GGCCAAGTCT GTACAACATC TTGAGTCCCT TTATACCTCT ATTACCAATT TTCTTTTGTC TTTGGGTATA CATTTGAACC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
841
#NZM-3 CTAATAAGAC TAAAAGATGG GGCTATTCCC TTAACTTCAT GGGCTATGTA ATTGGAAGTT GGGGTACCTT ACCACAAGAA CATATTATAC TCAAAATCAA AAACTGTTTT CGAAAACTTC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
961
#NZM-3 CTGTAAATAG GCCTATTCAT TGGAAGGTGT GTCAAAGAAT TGTGGGTCTT TTGGGATTTG CTGCCCCTTT TACACAATGT GGCTATCCTG CCCTAATGCC TTTGTATGCA TGTATTCAAG
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
1081

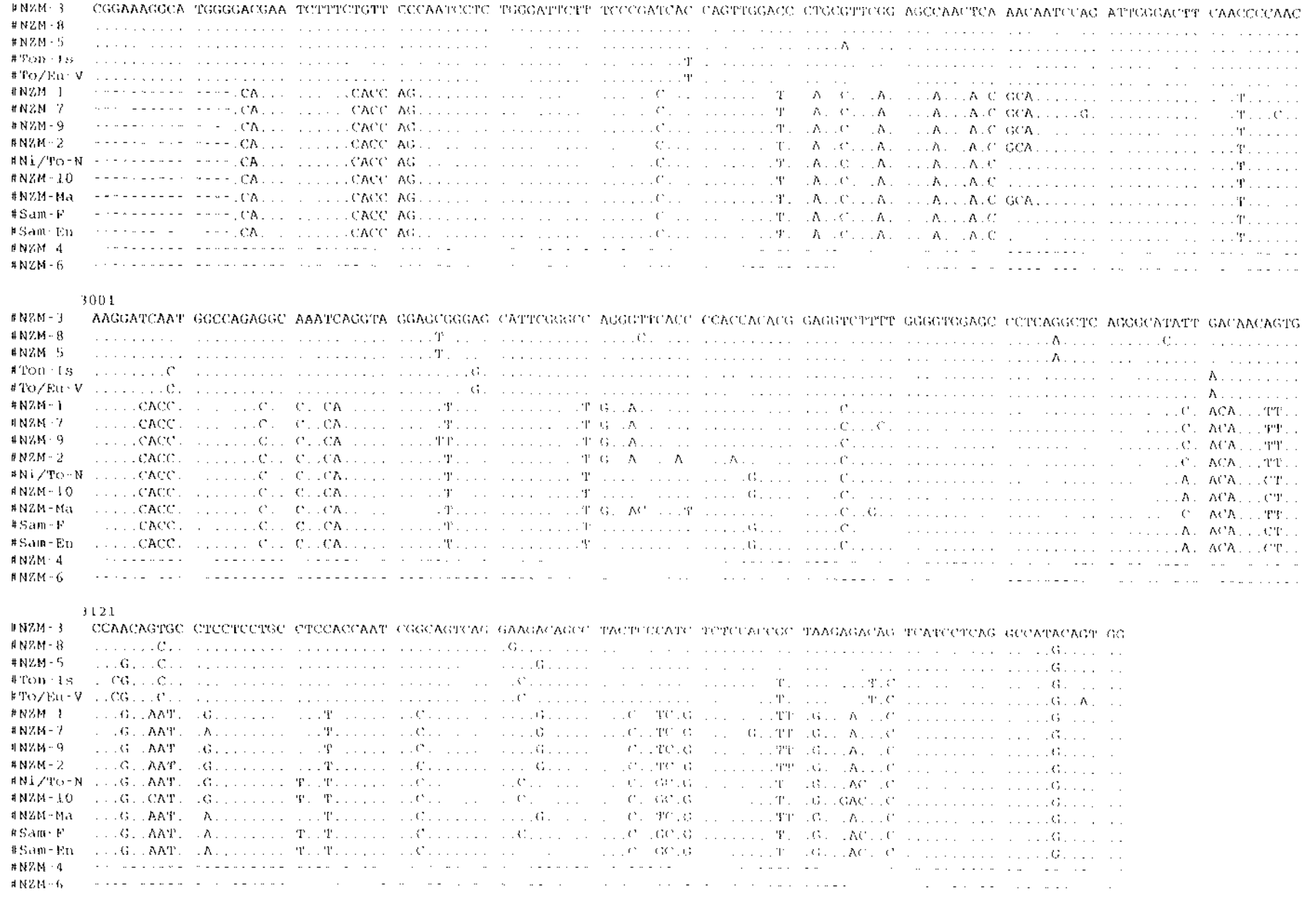



1441
#NZM-3 CGCTGAATCC CGCGGACGAC CCGTCTCGGG GCCGCTTGGG GATCTACCGT CCCCTCCTTC ATCTGCCGTT CCGGCCGACC ACGGGGCGCA CCTCTCTTTA CGCGGTCTCC CCGTCTGTGC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
1561
#NZM-3 CTTCTCATCT GCCGGACCGT GTGCACTTCG CTTCACCTCT GCACGTCGCA TGGAGACCAC CGTGAACGCC CACATGGTAT TGCCCAAGGT CTTGTATAAG AGGACTCTTG GACTCTCAGC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
1681
#NZM-3 GATGTCAACG ACCGACCTTG AGGCATACTT CAAAGACTGT GTGTTTAAAG ACTGGGAGGA GTCGGGGGAC GACATTAGGT TAATGATCTT TATACTAGGA GACTGTAGGC ATAAATTGGT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
1801


#NZM-3 CTGTTCACCA GCACCATGCA ACTTTTT CA CCTCTGCCTA ATCATCTCTT GTTCATGTCC TACTGTTCAA GCCTCCAAGC TGTGCCTTGG GTGGCTTTAG GACATGGACA TTGACCCTTA
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-Eo
#NZM-4
#NZM-6
1921
#NZM-3 TAAAGAATTT GGAGCTTCTG TGGAGTTACT CTCTTTTTTG CCTTCTGATT TCTTTCCATC TATTCGAGAC CTCCTCGACA CCGCCTCAGC TCTGTACAGG CAGGCCTTAG AGTCTCCGAA
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
2041
#NZM-3 ACATTGTTCA CCTCACCATA CAGCACTCAG GCAAGCTGTT CTGTCTTCGG GTGAGTTAAT CAATCTGGCT ACCTGGCTGG GAAGTAATTT GGAAGACCCA GCATCCAGGC AATTAGTGGT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
2161


#NZM 1 CAGTTATGTC AACATTAATA TGGGCCTAAA ACTCAGACAA CTATTGTGGT TTCACATTTC CTGTCTTACT TTTGGAAGAG AAACTGTTCT TGAATATTTG GTGTCTTTTG GAGTGTGGAT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-Eu
#NZM-4
#NZM-6
2281
#NZM-1 TCGCACTCCT ACCGCTTACA GACCACCAAA TGCCCCTATC TTATCAACAC TTCCGGAAAC TACTGTTGTT AGACGA--- ---CGAGGCAG GTCCCCTAGA AGAAGAACTC CCTCGCCTCG
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM 10
#NZM-Ma
#Sam-F
#Sam-Eu
#NZM-4
#NZM-6
2401
#NZM-1 CAGACGAAGG TCTCAATCAC CGCGTCGCAG AAGATCTCAA TCTCGGGAAT CTCAATGTTA ATATCCCTTG GACTCATAAG GTGCGAAAGT TTACTGGGCT TTATTCTTCC ACTGTACCTG
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM 7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-Eu
#NZM-4
#NZM-6
2521


#NZM-3 TCTTTAATCC TGAATGGCAA ACTCCCTCTT TTCCTGACAT TCATTTAAAA GAAGACATTA TCAATAGGTG TCAACAATAT GTGGGCCCTC TTACTGTTAA TGAAAAAAGA AGATTAAAAT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
2641
#NZM-3 TAATTATGCC TGCTAGGTTT TATCCTAACC TTACCAAATA TTTGCTTTTA GATAAAGGCA TTAAACTCTA TTATCCTGAA CACGCAGTTA ATCATTACTT TAAAACTAGG CATTATTTAC
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
2761
#NZM-3 ATACTCTGTG GAAAGCGGGC ATTCTATATA AGAGAGAAAC TACACGCAGC GCCTCATTTT ATGGGTTACC ATATTCTTCG GAACAAGAGC TACAGCATGG GAGGTTGGTC TTCCAAACCT
#NZM-8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6
2881


#NZM-3 CGGAAAGGCA TGGGGACGAA TCTTTCTGTT CCCAATCCTC TGGGATTCTT TCCCGATCAC CAGTTGGACC CTCCGTTCGG AGCCAACTCA AACAATCCAG ATTGGGACTT CAACCCCAAC
3001
#NZM-3 AAGGATCAAT GGCCAGAGGC AAATCAGGTA GGAGCGGGAG CATTCGRGCC AGGGTTCACC CCACCACACG GAGGTCTTTT GGGGTGGAGC CCTCAGGCTC AGGGCATATT GACAACAGTG
3121
#NZM-3 CCAACAGTGC CTCCTCCTGC CTCCACCAAT CGGCAGTCAG GAAGACAGCC TACTCCCATC TCTCCATCGC TAAGAGACAG TCATCCTCAG GCCATACAGT GG

3.1.2 Pacific HBV sequence features

Individual features of each Pacific sample sequence were investigated. When analysing a small sample set, it is necessary to compare differences against a larger database in order to get a true assessment of what is, or is not, unique. These samples were compared to the assembled database (see the later section on the HBV complete genome Genbank search), and the included disc file 'Database alignment'. Any mutations within the viral transcripts and transcriptional control elements (Table 1.1, 1.2 and Appendix A) were recorded. Where appropriate, the viral transcripts were translated to check whether the nucleotide changes affected the amino acid coded for, that is, if they were silent (synonymous) or replacement (nonsynonymous) mutations. The sequence variations noted in Table 3.4 are of amino acid changes that were completely unique, or nucleotide mutations that weren't necessarily unique but considered relevant to viral function (such as the 1896 stop codon which corresponds to HBeAg negative seroconversion) or deletions. Table 3.4 gives a general overview of the types and frequency of mutations that have occurred in the different Pacific samples.

The infected individuals were all asymptomatic for HBV infection, but had different states of liver dysfunction as measured by their alanine aminotransferase levels (ALT). The following is a description of the donors and the mutations found in the sequence. The implications of the results found here are discussed in chapter four.

3.1.2.1 Kawarau samples

The first group of ten samples is from the Whakatane child health, hepatitis foundation detected during a HBV survey of Kawarau.

NZM-1

Serum was extracted from an HBeAg positive, 21 year old Maori male with mild liver disfunction. Within this genome there were two stop codon mutations truncating separate transcripts. The first was in the Ribonuclease H enzyme transcript at polymerase (P) amino acid (aa) position 781, the second, in the X gene transcript at X protein amino acid 1065. The first stop codon was created by a single nucleotide change at bp 1430, the second, an 8 bp deletion from bps 1763-1770. The deletion within the X gene (which is believed to be a trans-acting activator), also overlaps with the core/pregenomic promoter and the enhancer II region. In addition, within the X gene there was a second nucleotide mutation that changed an arginine to a threonine at X gene amino acid position 946. A replacement though nonunique change also occurred in the polymerase gene at amino acid 371.

NZM-2

The serum was extracted from a Maori male of 19 years, who was HBeAg positive with mild liver disfunction. Mutations were found at bps 3059 and 3065 which lead to amino acid changes in the surface antigen, and the spacer region of the polymerase gene. They also effected the preS I and pre S II/S promoter regions. Moreover there were also changes in the core/pregenomic promoter region bp 1750 and a nonunique change in the X promoter at bp 1141.

NZM-3

Serum was extracted from a male Pacific Islander of 22 years with mild liver disfunction. This individual had the 1986 stop codon as well as the 1899 G to A mutation, and had been HBeAg negative for at least 11 years. In the core (bps 2038, & 2290), X protein (bp 1782), preSII & spacer regions (bp 3214), and terminal protein (bp 2820), there were a range of unique and common amino acid mutations. Also the enhancer I (bp 1053) and core/pregenomic promoter (bp 1782) sites contained nucleotide changes. The nucleotide mutation at the bp 3214 created an amino acid change in the initial methionine amino acid of the preS II gene.

NZM-4

Serum was extracted from a Maori male of 33 years he was HBeAg negative with mildly abnormal liver function. This individual came from a family with a history of HBV infection and cancer and was probably infected perinatally. Information was limited to bases 1-910 which cover the surface gene and parts of the polymerase. Within these regions mutations were found affecting both the polymerase/reverse transcriptase and the surface gene (bps 275, 392, 429,827). There was nothing identified in the sequence to explain why the entire genomes would not sequence.

NZM-5

Serum was extracted from a Maori male of 34 years who had recently seroconverted to HBeAg negative and had normal liver function. This individual had changes in the core/pregenomic promoter region (eg bps 1754,1766,1768), and in the X promoter (bp1220). They were not unique nucleotide changes and appeared repeatedly within the data base. Also, there were mutations in the SI promoter region (bp 2725) and an amino acid change in the preSII region (bp 124). Worthy of note was the unique mutation in "*e*", the pregenomic encapsulation signal (bp 1916), which could presumably affect mRNA pregenomic packaging however, this region is variable therefore may not be highly constrained. This mutation in the *e* corresponds with an amino acid change in the core gene.

NZM -6

Serum was taken from a 33 year old Maori male who was HBeAg negative with mildly abnormal liver function. No mutations were recorded in the portion of the genome that was sequenced, therefore no reason could be found as to why the rest of the genome could not be sequenced.

NZM -7

Serum was extracted from a Maori female of 34 years who was HBeAg negative and had seroconverted at least 3 years ago. This individual while asympotmatic had ALT levels of 198, and thus had definite liver disfunction. The viral genome demonstrated many mutations in all four genes (e.g. bps 16, 44, 302, 1616, 2247, 2304, 2341, 2342, 2451) as well as the preS II/surface, core/pregenomic, and X promoter regions, and the NRE (bp 3185, 3078, 1766, 1768, 1252). Some of these were unique, and others had dual effects eg bps 16, 44 and 1616. This genome had more mutations that the other samples.

NZM -8

Serum was extracted from an asymptomatic Maori male of 51 years who was HBeAg positive. This individual had mild levels of liver disfunction and was infected with both HBV-**C** and HBV-**D** genotypes. The genome was highly variable with common and unique amino acid variations found in the X promoter (bps 1116, 1118, and 1220) and X gene (bps 1635, 1636) as well as in the core (bps 2045, 2138, 2440, and 2452) and polymerase/reverse transcriptase gene (bp 843). Two unique nucleotide mutations within the X promoter corresponded to a change from valine to methionine within the polymerase/reverse transcriptase gene (bps 1116, & 1118).

NZM -9

Serum was extracted from male Maori of 34 years who was HBeAg negative with elevated ALT levels demonstrating definite liver disfunction. Interestingly, while this individual had been HBeAg negative for a prolonged (though unknown) period of time there were only a few mutations (bps 1655, 1754, 1768, 1770, 2354, 2371) with only one unique change in enhancer I (bp 1010).

NZM -10

Serum was extracted from a Maori male of 34 years who was HBeAg negative with mild liver disfunction. Unique mutations were found in the core (bp 2177), spacer (bp 3194), and polymerase/reverse transcriptase (bp 766) genes, as well as the S II promoter region (bp 3126). A few nonunique mutations were also present throughout the genome (eg bps 17, 2179, 2449). As well as the 1896 mutation this individual had the well reported 1899 G to A mutation.

3.1.2.2 Auckland samples

The following samples are from Pacific Island and Maori woman living in Auckland found to be HBV positive by routine testing during pregnancy. These women all had normal liver function.

NZM-Ma

Serum was extracted from Maori female who was HBeAg positive. This genome had a 33 bp deletion in the core gene (bps 2158-2190) and an 8 bp deletion in the X gene (bps 1763-1770) that created a stop codon (bps 1785-1787) (truncating the protein further). This deletion, as in NZM-1, also mutated the core/pregenomic promoter and the enhancer II region.

Ni/To-N

Serum was extracted from a part Niuean/part Tongan female who was HBeAg positive. This sequence was wild type with no mutations.

Sam-En

Serum was extracted from a Samoan female who was HBeAg positive. With only four mutations throughout, this was a relatively conserved genome. The mutations that did occur were within the X and surface gene (bps 1559, 1724, 383) with one in the enhancer II region (bp 1724). Interestingly this genome has the 1896 stop codon mutation but remains HBeAg positive.

Sam-F

Serum was extracted from a HBeAg positive Samoan female with no detectable mutations, and the predominant sequence was wild-type.

To/En-V

Serum was extracted from a Tongan female, who was HBeAg positive. This sequences had only a few mutations, two in the X promoter region that were nonunique and one each in the in the X gene (bp 1679) and spacer region (bp 3218) that were unique.

Ton-Is

The donating individual was a Tongan individual who was HBeAg positive. Mutations occurred in the preS I promoter TATA box with an A to G substitution (bp 2785), and in both enhancer regions all four of which were unique (bps 970, 977, 1714, 1717).

These genomic sequences display a range of mutations that are randomly distributed in number and position. The implications of these genomic variations are discussed in Chapter 4

Table 3.4 The Pacific hepatitis B virus samples', sequence features, highlighting the various mutations

	NZM-1	NZM-2	NZM-3	NZM-4	NZM-5	NZM-6	NZM-7	NZM-8	NZM-9	NZM-10	NZM-Ma	Ni/To-N	Sam-En	Sam-F	To/En-V	Ton-Is
1896 Stop codon			*		*		*		*	*			*			
Viral Transcript																
pre S I		**														
pre S II			*		*		*									
surface				*			*						*			
precore																
core			*		*		*	***		**	*33**					
polymerase																
terminal protein			*				*									
spacer		**					***			*					*	
DNA polymerase/reverse transcriptase				*			**	**		*						
ribonuclease H	*															
X protein	**8*		*				*				*8*		**		*	
Viral transcript control elements																
pre S I promoter					*											*
pre S II/S promoter		**					*			**						
core/pregenomic promoter	*8	*	*								*8					
X promoter								**								
enhancer I									*							**
enhancer II													*			**
direct repeat I																
direct repeat II																
e					*											
negative regulatory element							*									

Key.
Blue= Has 1896 stop codon mutation. Grey= region not sequenced.
Green= genotype HBV-**C.** * = stop codon mutation.
* = start codon mutation. * = deletion mutation, numbers beside indicate number of deleted bases. * = unique mutation.

3.1.3 Pacific HBV phylogenetic analysis results

Analysis of the 16 samples was conducted using sites 1—910 (as this was the maximum available for all sixteen). This region contained an overlap of the polymerase and surface genes and was therefore comparatively conserved, yet was sufficiently variable to give excellent resolution of the samples.

Pairwise distance (observed and corrected), parsimony and maximum likelihood criteria were all used to determine the relationships within the Pacific samples. Bootstrap values were obtained with each of the above criteria to give a measure of tree stability (Table 3.5). In addition, the bootstrap values from 1000 replicates of observed distances were further analysed using the program NewBoots (Table 3.6). Nearest-neighbor bootstrap (NNB) values give a measure of local tree stability.

For all criteria and search strategies, the samples consistently separated into two genotypes HBV-**C** and HBV-**D** (Figure 3.3), with each genotype further dividing into two sub-groups (Figures 3.2. and 3.3) With all criteria the bootstrap values strongly support these four splits (Tables 3.5 and 3.6)

HBV-**C** (NZM-3, NZM-4, NZM5, NZM-8), (Ton-Is, To/Eu-V)

HBV-**D** (NZM-1, NZM-2, NZM-6, NZM-7, NZM-9), (Sam-F, Sam-En, NZM-Ma, NZM-10)

The overall tree is stable giving a valid estimate of the phylogenetic relationships. The co-infection of genotypes HBV-**C** and HBV-**D** in sample NZM-8 does not affect the underlying tree. This sequence groups with NZM-3,4, and 5 (see above). While the NZM-3,4,5,8 cluster has a lower boot strap support value (averaging at 93%), the NNB value for this cluster is 100% and removal of NZM-8 from the tree gives a bootstrap value of 100% for the NZM-3,4,5 group.

This analysis demonstrates a definite split between the Pacific-specific HBV-**C** genomes, separating New Zealand Maori from the Tongan sequences. The HBV-**D** genomes appear intermixed, with Samoan and NZ Maori genomes clustering together. Overall, the 16 Pacific samples clearly divided into two genotypes, HBV-**C** and HBV-**D**. To confirm the genotype of these Pacific samples and placement on a global scale, comparison against a worldwide distribution of different genotypes is necessary. This is the content of 3.2 HBV complete genome Genbank search .

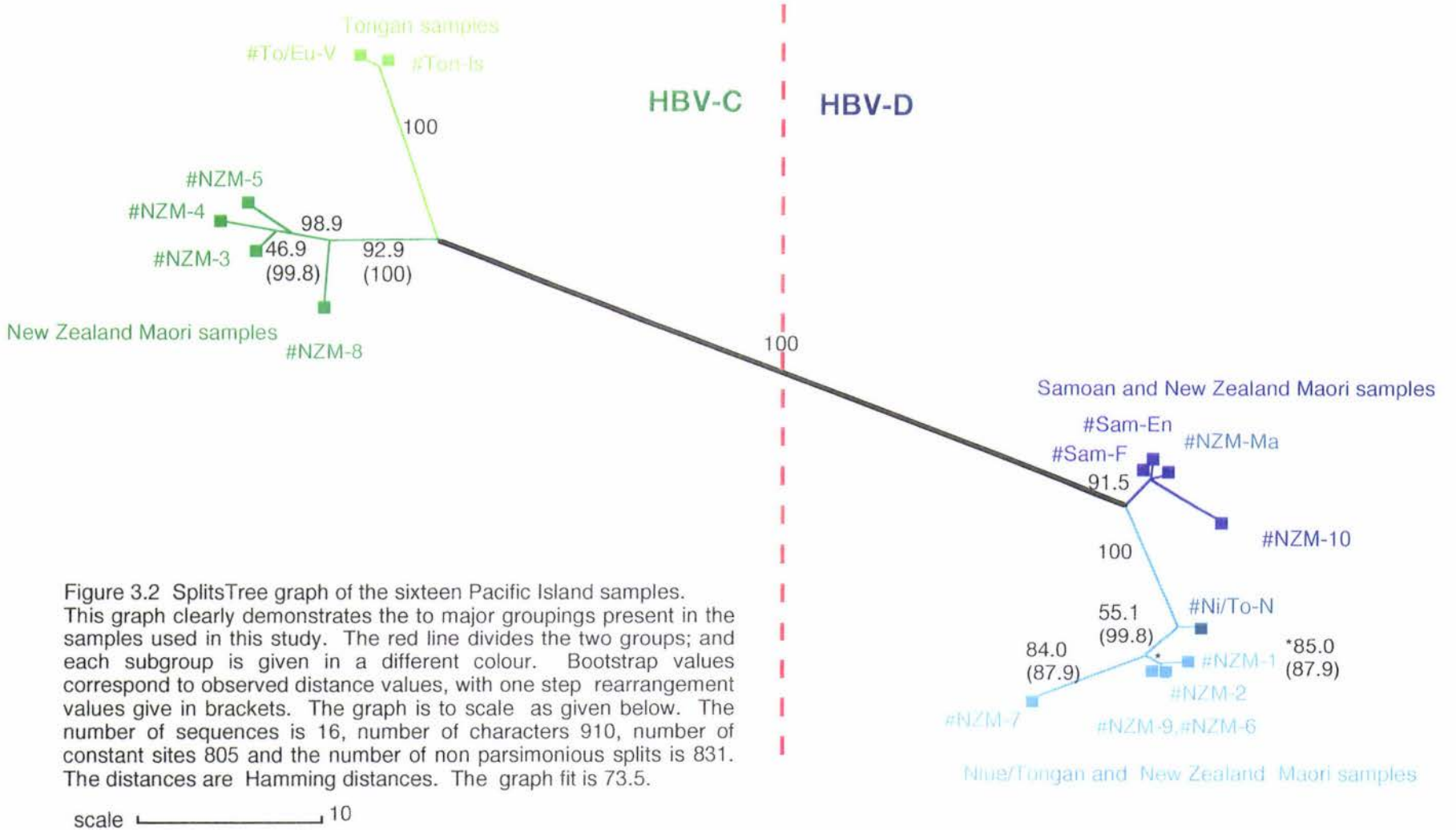


Figure 3.2 SplitsTree graph of the sixteen Pacific Island samples. This graph clearly demonstrates the to major groupings present in the samples used in this study. The red line divides the two groups; and each subgroup is given in a different colour. Bootstrap values correspond to observed distance values, with one step rearrangement values give in brackets. The graph is to scale as given below. The number of sequences is 16, number of characters 910, number of constant sites 805 and the number of non parsimonious splits is 831. The distances are Hamming distances. The graph fit is 73.5.

scale ——— 10

Table 3.5 Bootstrap values for branches for the Pacific hepatitis B virus samples' clusters (sites 1-910)

Sets included in the consensus tree

1 1234567890 123456	Cluster	Phylip obs.dis. 1000	PAUP* cor.dis. 100	PAUP* Parsimony 100	Splits Hamming 100	Branch length NJ
.********* *.....	(3,4,5,8,10,Ma,En,F,V,Is)	100.00	100.00	100.00	100.00	13.65
.....***** *.....	(3,4,5,8,V,Is)	100.00	100.00	100.00	100.00	57.33
.........* *.....	(V,Is)	100.00	100.00	100.00	100.00	13.65
.....***..	(3,4,5)*	98.90	98.00	86.00	83.00	5.46
.....****.	(3,4,5,8)	92.90	82.88	96.00	100.00	5.46
.****.....	(10,Ma,En,F)	91.50	63.00	86.33	96.00	0.91
.********* ***...	(10,Ma,En,F,3,4,5,8,V,Is,7,N)	85.30	70.00	30.00		
.********* *.*...	(10,Ma,En,F,3,4,5,8,V,Is,N)	55.10	71.00	68.00	86.00	1.82
.....**...	(3,4)	46.90	53.00	29.28		
.***.....	(10,Ma,En)	40.40	42.00			
..**.....	(Ma,En)	32.00	34.00			
********** ***.*.	(10,Ma,En,F,3,4,5,8,V,Is,7,N,9,2)	27.60	18.00			
.********* ***.*.	(10,Ma,En,F,3,4,5,8,V,Is,7,N,9)	26.30	19.00			

Samples in order:

1 #NZM-2
2 #NZM-10
3 #NZM-Ma
4 #Sam-En
5 #Sam-F
6 #NZM-3
7 #NZM-4
8 #NZM-5
9 #NZM-8
10 #To/Eu-V
11 #Ton-Is
12 #NZM-7
13 #Ni/To-N
14 #NZM-1
15 #NZM-9
16 #NZM-6

Key

- Each row represents a split/cluster in the tree, and is demonstrated by presence in, by "*" or absence from, by ".".
- The numbers in the first column correspond to the taxa as listed beneath it.
- The 'Cluster' column names the samples included in the split.
- Highlighted are the genotype and subgroup clusters.
- The Branch length NJ (Neighbour-joining) column gives the number of nucleotide bases creating the split.

* This split has a bootstrap value of 100% if NZM-8 is excluded from the data set.

- The obs.dis. column gives the observed distances where as the cor.dis. column gives the corrected distances

These bootstrap values show clear agreement — any discrepancies lie in the different resolving power of the methods of analysis. The Maximum Likelihood trees (with out bootstraps) give the same clusters. Using complete genomes, does not change the tree (see Appendix C).

Table 3.6 Nearest neighbour bootstrap one step rearrangement values from one thousand replicates of the observed distances from the Pacific hepatitis B virus samples tree.

Sub group split	Standard bootstrap	rearrangement		total1 step
		a	b	
1	100.0	0.0	0.0	100.0
2	100.0	0.0	0.0	100.0
3	100.0	0.0	0.0	100.0
4*	98.9	0.0	0.0	98.9
5	92.9	7.1	0.0	100.0
6	91.5	0.0	0.0	91.5
7	85.3	1.1	0.0	86.4
8	55.1	44.7	0.0	99.8
9	46.9	41.6	11.3	99.8
10	40.4	37.10	3.8	81.30
11	32.00	26.2	24.4	82.6
12	27.6	25.40	5.4	58.4
13	26.30	26.2	7.0	59.50

Number of partitions 50

These NNB values demonstrate the local and, in this case overall, stability of the Pacific samples tree. In Table 3.6 the second columns are the standard bootstrap values from observed distances. Columns three and four are the percentage values for the two rearrangements for each internal edge (branch), that is, the two rearrangements one-step from the consensus. The last column is the sum of the previous three columns giving the total one step rearrangement value. Sub group split 4* corresponds to the HBV-**C**, (NZM-3,4,5, and 8) cluster. The data shows that if a rearrangement around one partition is allowed, then the additive support sums to 100%. The lesser bootstrap values for rows 7 to 13 represent the resolution of the samples within the genotype clusters. Here consistent branching within the clusters could not be determined; for example there was 55.1% support for the cluster NZM-1,2,6,7 and 9, and 44.7% for the conflicting split NZM-2,6,7,10, Ni/To-N. However, the additive NNB value of these two splits sums to 99.8% demonstrating that despite the lack of resolution within the clusters they consistently grouped together in specific group clusters — that is the tree is locally stable around this edge.

3.2 Hepadnaviridae complete genome, Genbank search

A total of 80 complete human HBV sequences were retrieved from Genbank together with hepatitis sequences from gibbon, chimpanzee, and woodchuck. Background information including the accession numbers, references and geographical origin are provided in appendix B. Their sequence alignment is recorded on disk, in a Se-Al file called "Database alignment" (the Se-Al program is also included on disc). A duck hepatitis B viral sequence was also obtained but as it was greatly divergent from both the wood chuck and primate strains that alignment proved to be too difficult and consequently it is not included in the database. The sequences are numerous and demonstrate an extensive array of variations. While all sequences are included in the database alignment, only a selected sample were used in phylogenetic analysis (Table 3.7). Criteria for inclusion depended on the frequency of genotypes, for example, there were only two HBV-**E** and three HBV-**F** genomes on Genbank at the time of collection, so all five were included. HBV-**B**, -**C**, -**D** genomes were numerous and therefore were selected for a wide geographical distribution, and preferably lack of the 1896 stop codon (indicating a potential wild-type genome).

3.2.1 Database and Pacific samples phylogenetic analysis

The 14 completely sequenced genomes of the Pacific samples were combined with complete genome sequences of 24 other human, a gibbon, a chimp a woolly monkey and 4 wood chuck sequences.

Using parsimony, maximum likelihood, neighbour-joining and spectral analysis the placement of the Pacific samples within the database was determined (Figure 3.3). The two groups unequivocally divided into genotypes HBV-**C** and HBV-**D**. The Pacific samples that were of genotype HBV-**C** consistently cluster with the Pacific strains, taxa 18 and 19 of French Polynesia and New Caledonia respectively. The HBV-**D** samples divided giving two separate clusters, a deep divergence outside of the other HBV-**D** strains and a more recent branching within the HBV-**D** strains. This is not clearly shown in figure 3.3 the SplitsTree graph, however, it is in the maximum likelihood trees (see Appendix C). The trees determined by each of the different criteria show identical branching structure between the genotypes and are supported by strong bootstrap values (Table 3.8). Branching within the genotype clusters is not as stable, but there is no variation between the criteria in placement of the Pacific sample

Table 3.7 Hepadnaviridae sequences obtained from the Genbank database used in the phylogenetic analysis. The taxa column gives the numbers associated to that sequences in all analysis. The country of origin, genotype, Genbank accession number, serotype (where known), and the clinical of the donor are all listed as well as the reference paper giving the place of virus infection. The GT column refers to the genotype

Taxa	Name	GT	Accession #	Subtype	Clinical state	Reference
1	Colombia	F	X75663	adw4q-	not given	Norder *et al.* (1994)
2	Brazil	F	X69798	adw4	not given	Naumann *et al.* (19930
3	France 2	F	X75658	adw4q-	carrier	Norder *et al.* (1994)
4	Africa	E	X75657	ayw4	not given	Norder *et al.* (1994)
5	Africa 2	E	X75664	ayw4	carrier	Norder *et al.* (1994)
6	Germany 1	A	X70185	adw2.	HBeAg-positive infection	Preisler-Adams (1993)
7	Polish 2	A	Z35717	adw2	not given	Plucienniczak (1994)
8	U.S.A. 3	A	S50225	adw2	persistent hepatitis	Wands *et al.* (1992)
9	German 7	A	Z72478		not given	unpublished
10	Asian 12	B	X98077	adw	carrier	Pult *et al.* (1997)
11	Japan 14	B	D23677		carriers	Horikita *et al.* (1994)
12	Japan 10	B	D00329	adw	carrier	Okamota *et al.* (1988)
13	Indonesia	B	D00331	adw	carrier	Pult *et al.* 1997
14	China 6	B	X97851	adw	carrier	Alexpopulou *et al.* (1996)
15	Japan 20	C	D50517		chronic hepatitis	Asahina *et al.* (1996b)
16	Japan 7*	C	D28880		carrier	Moriyama *et al.* (1994)
17	Japan 6	C	L08805	adr	carrier	Ogata *et al* . (1993)
18	F/Polynesia	C	X75656	adrq-	not given	Norder *et al.* (1994)
19	N/Caledonia	C	X75665	adrq-	carrier	Norder *et al.* (1994)
30	Greece 1	D	X97848	ayw	carrier who developed fatal HBV	Alexpopulou *et al* .(1996)
31	Greece 6	D	X80925	ayw	carrier	Karayiannis *et al.* (1995)
32	Germany 2	D	Y07587	adw	immunocompromised child	Stoll-Becker *et al.* (1997)
33	Israel	D	L27106		fatal fulminant	Hasegawa (1994)
38	gibbon		U46935	ayw 3g	chronic hepatitis	Norder *et al.* (1996)
39	chimp		D00220	LSH	"naturally" infected chimp	Vaudin *et al.* (1988)
40	woolly monkey		AF046996		"naturally" infected w. monkey	Lanford *et al.* (1998)
41	WHV		M18752	WHV 7	wood chuck	Cohen et al. (1988)
42	WHB		J0244		wood chuck	Galibert et al. (1982)
43	WHV.		M19183	WHV 8	wood chuck	Cohen et al . (1988)
44	_WHV_		J04514		wood chuck	Giranes et al. (1989)

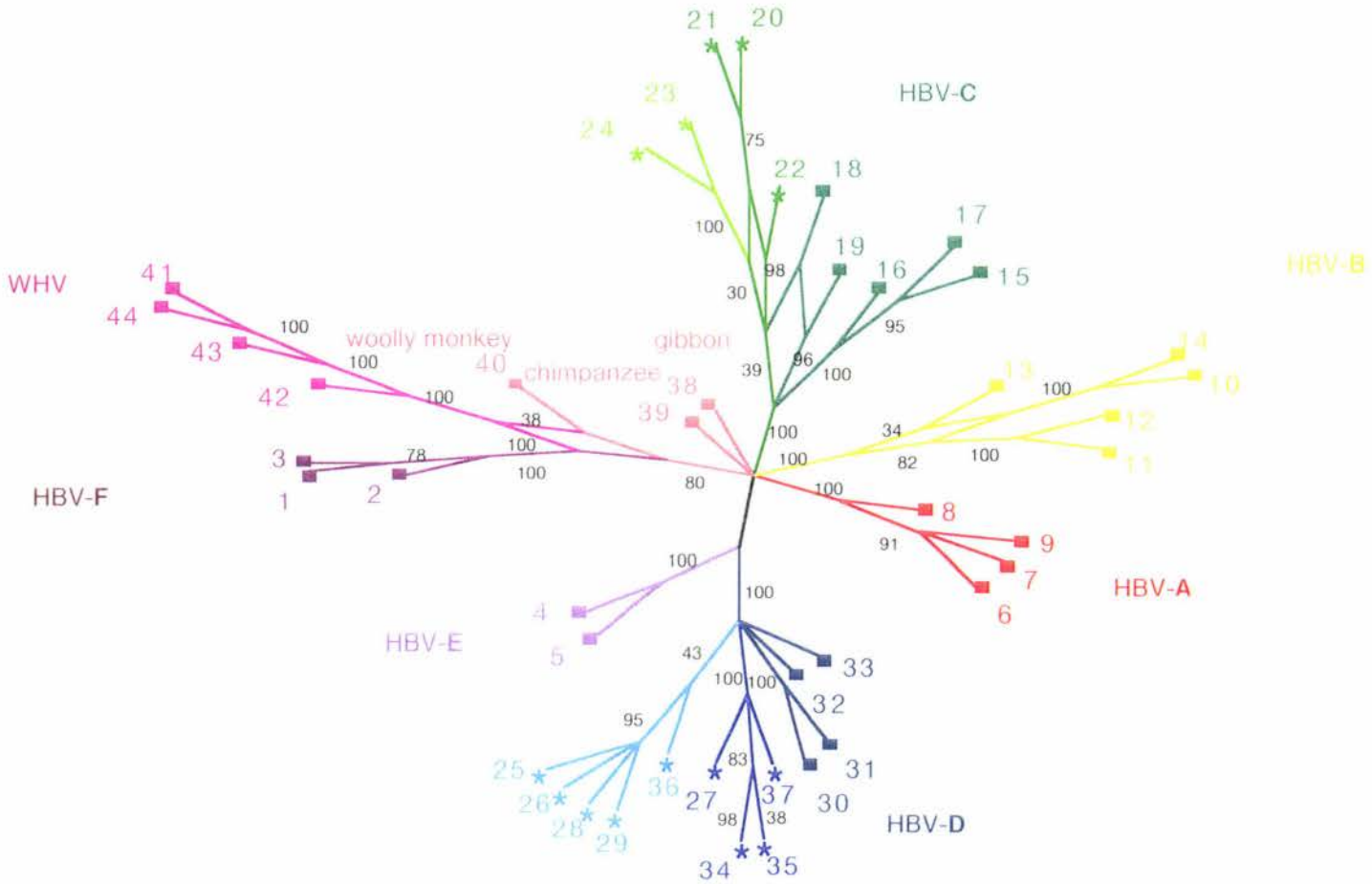


Figure 3.3 SplitsTree graph showing phylogenetic network of the database and Pacific hepadnaviruses. The numbers correspond to the sequences as given in Table 3.9. Taxa marked with * are Pacific samples sequences. The lighter and darker colours of blue and green represent the different subgroups of the Pacific samples as shown in Figure 3.2. Bootstrap values are based on observed Hamming distances. The Pacific samples cluster within the HBV-C (green) and the HBV-D (blue) strians. This graph is not to scale .

Fit=74.5 ntax=44 nchar=3058 (of 3390) -bootstrap 100 -dsplits -hamming

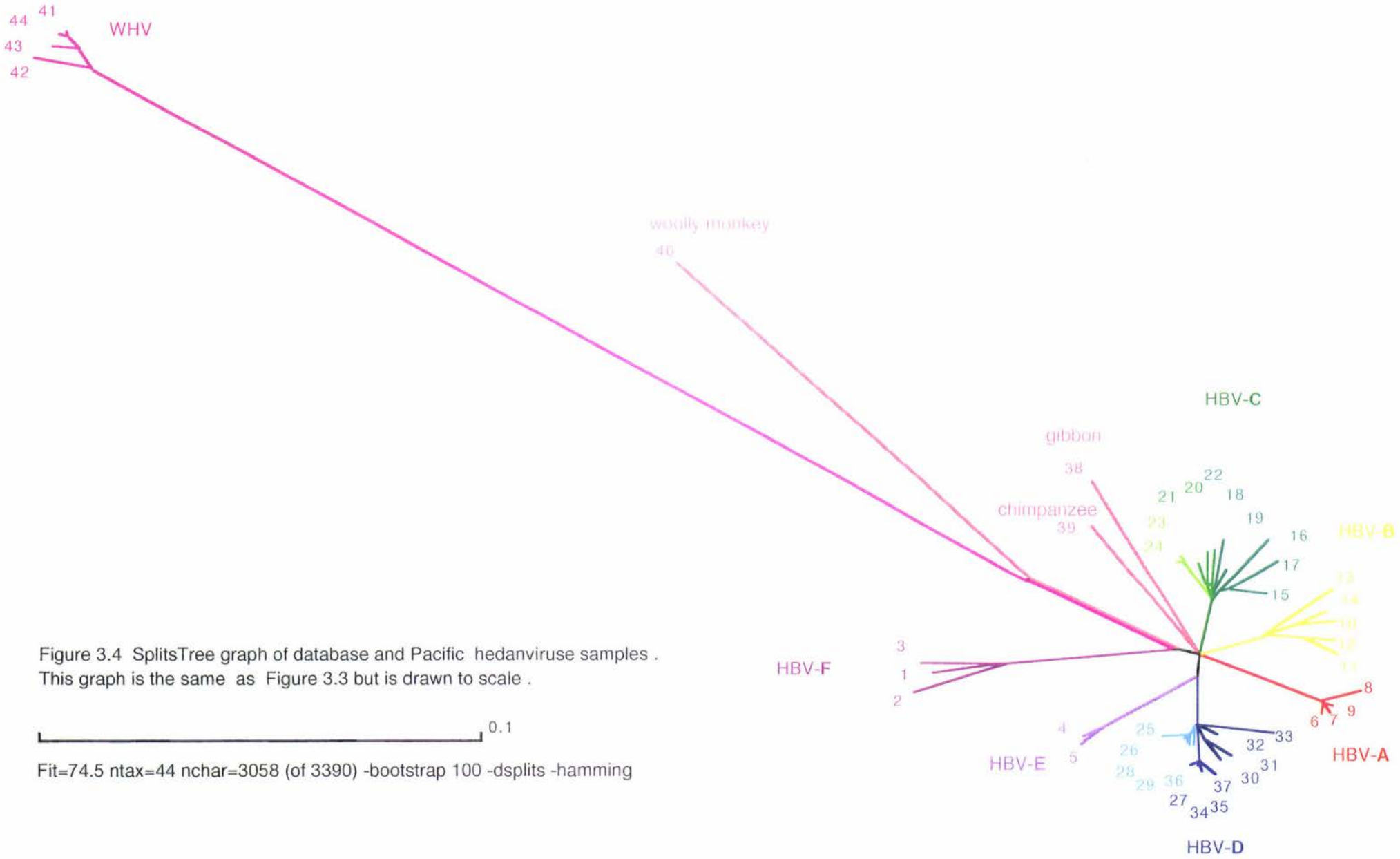


Figure 3.4 SplitsTree graph of database and Pacific hedanviruse samples . This graph is the same as Figure 3.3 but is drawn to scale .

Fit=74.5 ntax=44 nchar=3058 (of 3390) -bootstrap 100 -dsplits -hamming

Table 3.8 Bootstrap values form the combined database and Pacific hepadnavirus samples tree. The first row gives the number for the corresponding sequence given in Table 3.9. The bootstrap values are from Neighbour-joining corrected distances and SplitsTree observed Hamming distances; "*" indicates presence in a cluster and "." absence from cluster.

1 2 3 4 4	NJ	Splits
12345678901234567890123456789012345678901234	Cor/dis	Obs/dis
.....*...*.....	100.00	100.00
.....****	100.00	100.00
.....**	100.00	75.00
.....**	100.00	100.00
.....*.....**.*.....	100.00	100.00
.....**	100.00	98.00
.....****	100.00	100.00
.....*****	100.00	100.00
.....***	100.00	70.00
.....**.**.....	100.00	95.00
.....**	100.00	100.00
.....**********	100.00	100.00
...**.....	100.00	100.00
.....**.*.....	99.00	*25.00*
...**********	99.00	100.00
.....**********	99.00	100.00
...**.....**********	99.00	90.00
.....***.*.....	96.00	82.00
.....*.*.....	96.00	95.00
.....**.******..*.....	96.00	*24.00*
.....**	92.00	Ω
.....**	91.00	83.00
.*.**********	90.00	78.00
.....*...*.....	89.00	*1.00*
.....**.*.....	88.00	91.00
.....*.*.....	85.00	Ω
.....*****	81.00	30.00
...**********	79.00	80.00
.....**	78.00	96.00
.....*******	74.00	Ω
.....**********	68.00	Ω
.....**	66.00	80.00
.....**.**..**.....	60.00	Ω
.....<.....**.**..**..*.....	58.00	Ω
.....*****	55.00	100.00
.....**.**.....	53.00	96.00
...**********	52.00	80.00
.....**.....	47.00	18.00
.....*.**	47.00	100.00
.....***.....	46.00	98.00
.....**********	42.00	Ω
...**********	40.00	38.00
.....***	40.00	100.00
.....**.******.....	40.00	Ω
.....********.....	35.00	Ω
.....**.**.....*.....	34.00	43.00
.....**.	33.00	Ω
.....*..*	28.00	100.00
.....*..*.....	26.00	Ω
.....**	23.00	5.00
.....*.*.....	22.00	Ω
.....**********.....**.....	20.00	Ω
.....****.....********.....	15.00	Ω
.....*.*****.....	15.00	39.00
...**********.*.....	14.00	Ω
.....*..**.....	12.00	Ω
.....*.*	12.00	Ω

KEY *red* = Indicates that these bootstrap values were not in the consensus SplitsTree graph.
Ω = Indicates where there are no corresponding splits in SplitsTree for that partition.

The neighbour-joining corrected bootstrap distances were stopped at 10.00% for Table 3.8. Corrected distances were used in neighbour-joining analysis and observed distances in the SplitsTree spectral analysis, these methods are not entirely comparable as they have different resolving ability and differ by two variables; corrected verses observed distances, and spectral analysis verses neighbour-joining. This is reflected in the variations in the splits represented. The bootstrap values not represented in the SplitsTree graph that are present in the Neighbour-joining values are indicated by the Ω .

The bootstrap values in Table 3.8 demonstrate that the genotype clusters and the major sub-groupings with in the tree are stable, supported by reasonable bootstrap values in both methods of analysis.

Table 3.9 Hepadnaviridae sequence numbers
Each sequence was assigned a number for clarity in reading the splits graphs.

[1] Colombia	[12] Japan-10	[23] #To/Eu-V	[34] #Sam-En
[2] Brazil	[13] Indonesia	[24] #Ton-Is	[35] #Sam-F
[3] France-2	[14] China-6	[25] #NZM-7	[36] #NZM-Ma
[4] Africa	[15] Japan20	[26] #NZM-2	[37] #NZM-10
[5] Africa-2	[16] Japan-7	[27] #Ni/To-N	[38] gibbon
[6] Germany-1	[17] Japan-6	[28] #NZM-1	[39] chimpanzee
[7] Poland-2	[18] F/Polynesia	[29] #NZM-9	[40] woolly monkey
[8] U.S.A.-3	[19] N/Caledonia	[30] Greece-1	[41] WHV
[9] Germany-7	[20] #NZM-3	[31] Greece-6	[42] WHB
[10] Asia-12	[21] #NZM-5	[32] Germany-2	[43] WHV.
[11] Japan-14	[22] #NZM-8	[33] Israel	[44] _WHV_

3.2.2 Covarion structure analysis

In the final paragraphs of section 1.2.2, the lack of resolution of the viral phylogeny was discussed. The long branch leading to the HBV-**F** genotype was noted as a key potential cause of this, with three hypotheses stated (genuinely older, faster rates, change in covarion structure). While it was not within the scope or aims of this thesis to sequence more HBV-**F** genotypes, nor was it possible to determine a rate of evolution for individual genotypes, it was however within the scope of this thesis to investigate the potential variation in covarion structure. Using the Inequality test of Lockhart *et al.* (1998) any change in covarion structure between the genotypes was investigated.

The test involves first calculating the proportion of each type of site in the individual genotype groupings (refer section 2.7.6 and Table 3.10). The significance cut off value of $Z \leq 2$ (pers. comm. Peter Lockhart, IMBS, Massey University) was used.

No apparent change in individual evolutionary structure between genotypes was observed. Consequently, the high number of changes cannot be explained by a change in covarion structure on this lineage. Note that a change in covarion structure cannot be excluded by this test (the power of the test is not known). It is just that the test cannot find any evidence for a change in covarion structure. Therefore, at present, either an older date or faster rate is more likely explanations for genotype **F** being the most divergent.

Table 3.10 Covarion Inequality test values from the hepatitis B virus genotypes' analysis.

Genotype combination	N1	N2	N1+N2	N3	N4	N5	X	S	Z
A/B	2739	139	2878	61	265	37	145.95	29.84	4.89
C/A	2658	111	2769	55	364	24	14.46	33.47	0.43
C/B	2587	65	2652	168	277	107	74.04	33.41	2.22
D/A	2593	160	2753	58	314	16	85.51	31.43	2.72
D/B	2524	105	2629	182	247	83	55.16	33.3	1.66
D/C	2467	80	2547	261	217	113	73.67	33.78	2.18
E/A	2824	267	3091	76	42	3	1.89	20.75	0.09
E/B	2702	196	2898	261	31	14	10.14	29.8	0.34
E/C	2640	156	2796	368	25	20	14.56	32.63	0.45
E/D	2681	103	2784	312	27	18	13.27	31.13	0.43
F/A	2641	292	2933	69	203	10	4.77	29.34	0.16
F/B	2561	218	2779	215	153	60	41.74	32.29	1.3
F/C	2495	188	2683	316	141	72	46.27	34.1	1.36
F/D	2470	200	2670	264	141	66	44.25	32.94	1.34
F/E	2679	283	2962	37	205	8	5.01	27.93	0.18

3.3 HLA-DPA1 data

Fifty-one samples of HLA-DPA 1 exon 2 were sequenced and typed. Their sequences are given in Table 3.12 with allele typing and frequency given in Table 3.11. Further results and analysis are written up as a manuscript in Appendix D that has been submitted to Tissue Antigens.

Table 3.11 Trobriand Islander HLA-DPA1 allele typing results.

HLA-DPA1 allele	Number
DPA1*01013	3
DPA1*01013/02022	4
DPA1*02022	44

This allele typing was a pilot study to investigate the viability of sequence based typing, the results demonstrated that the technique is very successful at determining heterogeneic sequences with greater resolution power than other contemporary techniques. It is interesting that there is such low diversity within the Trobriand Islanders. This, together with the variant that occur there (DPA 01013 and DPA 02022) are consistent with these populations being ancestral to the Polynesian people (see also Heagleburg *et al.* 1999a).

Table 3.12 Trobriand Islander HLA-DPA1 allele sequences

```
     1                                                                                                                  100
#010 GCGGACCATG TGTCAACTTA TGCCATGTTT GTACAGACCC ATAGACCAAC AGGAGAGTTT ATGTTTGAAT TTGATGAAGA TGAGCAGTTC TATGTGGATC
#017 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#064 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#079 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#080 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#081 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#083 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#084 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#085 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#086 ..........RY........ .....S.... .......... ...R...... .......... .......... ....MW.... .......... ..........
#087 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#088 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#089 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#090 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#092 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#093 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#094 ..........GC........ .....G.... .......... ...G...... .......... .......... ....AT.... .......... ..........
#095 ..........GC........ .....G.... .......... ...G...... .......... .......... ....AT.... .......... ..........
#096 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#098 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#099 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#100 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#101 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#102 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#105 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#106 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#107 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#108 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#110 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#111 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#112 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#113 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#115 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#118 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#119 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#120 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#122 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#127 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#131 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#132 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#133 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#134 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#137 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#138 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#141 ..........RY........ .....S.... .......... ...R...... .......... .......... ....MW.... .......... ..........
#142 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#143 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#146 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#147 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#148 ..........RY........ .....S.... .......... ...D...... .......... .......... ....MW.... .......... ..........
#149 ..........RY........ .....G.... .......... ...R...... .......... .......... ....MW.... .......... ..........
```


101
200
#010 TGGACAAGAA GGAGACCGTC TGGCATCTGG AGGAGTTTGC CCGAGCCTTT TCCTTTGAGG CTCATGGCGG GCTGGCTAAC ATTGCTATAT TGAACAACAA
#017
#064
#079
#080
#081
#083
#084
#085
#086 R
#087
#088
#089
#090
#092
#093
#094 A
#095 A
#096
#098
#099
#100
#101
#102
#105
#106
#107
#108
#110
#111
#112
#113
#115
#118
#119
#120
#122
#127
#131
#132
#133
#134
#137
#138
#141 R
#142
#143
#146
#147
#148 R
#149 R

```
    201                                                                 247
#010 CTTGAATACC TTGATCCAGC GTTCCAACCA CACTCAGGCC GCCAATG
#017 ....... ....... ....... ....... ....... ....... .
#064 ..... ... ....... ....... ....... ....... ..... .
#079 ..... ... ....... ....... ....... ....... ..... .
#080 ....... ....... ....... ....... ....... ....... .
#081 ....... ....... ....... ....... ....... ....... .
#083 ....... ... ....... ....... ....... ....... ..... .
#084 .. ..... ....... .. ....... ....... ....... ... .
#085 ....... ....... ....... ....... ....... ....... .
#086 ....... ....... ....... ....... ....... R...Y..
#087 ....... ....... ....... ....... ....... ....... .
#088 ....... .... ..... ....... ..... ....... ... .
#089 ....... ..... .. ....... ..... ....... .. .
#090 ....... ....... ....... ..... ....... ..... .
#092 ... .... ..... ... ....... .. ....... ..... .
#093 ....... ....... ....... ....... ....... ... .
#094 ....... ..... . ....... ..... ....... A..  C.
#095 ....... .....  ....... ..... ....... A..  C.
#096 ....... ....... ....... ..... ....... ..... .
#098 ....... ....... ....... ..... ..... ..... .
#099 ....... ....... ....... ..... ....... ..... .
#100 ....... ....... ....... .. ....... ..... .
#101 ....... ....... ..... ... ....... ..... .
#102 ....... ....... ....... ....... ..... .
#105 ....... ....... ....... ....... ..... .
#106 ....... ..... .. ....... ....... ..... .
#107 ....... .. .... ... .. ... ....... .. .
#108 ....... ....... . ....... .. ..... ..... .
#110 ... ..... ....... ....... - ....... ..... .
#111 ....... ....... ....... ....... ....... ..... .
#112 ....... ....... ....... ....... ....... ... .
#113 ....... ....... . ....... ....... ..... .
#115 ..... .... ....... ....... ....... ..... .
#118 ....... ....... ....... ....... ....... .. ... .
#119 ... ..... ....... ....... ..... ....... ..... .
#120 ....... ....... ....... ....... ..... ..... .
#122 ....... ....... ....... ....... .. ..... ... .
#127 ....... ....... ....... ....... ..... ... .
#131 ..... ... ....... ... ... ..... . ..... .
#132 ....... ....... ... ... ... .. .. .... .
#133 .. ..... ....... ....... ..... .. . .. ..... .
#134 ....... ....... ....... ....... ..... .
#137 ....... ....... ....... ... .. ..... ..... .
#138 ....... .. ..... ..... ... ..... .. ..... .
#141 . ..... ....... ..... ... ..... . R...Y
#142 .....  ..... ....... ..... ... . ..... .
#143 .....  ..... ... ....... ..... .. ... .
#146 ....... ..... ... ....... ..... . ..... .
#147 ... ..... ....... ....... ..... ..... .
#148 ..... ... ....... ..... .. ..... R...Y.
#149 ....... ....... ....... ..... R  .Y.
```

Chapter Four: Discussion and Conclusions

4.1 Pacific HBV samples sequence features

Fourteen complete and two partial HBV genomes from 16 asymptomatic HBV carriers were studied. Here, features of the genomes are discussed and related to the clinical states of the donors. It is important to understand that there is contradictory data on the clinical significance of the DNA mutations (Karayiannis 1995; Asahina *et al.* 1996a; Miyakawa *et al.* 1997) so my data cannot yet be interpreted definitively in these terms. It is also important to recall that the sequences obtained here are the predominant sequences, that there is a quasi-species distribution and consequently many other sequences are present in the host. Given also that there can be both positive and negative interactions between these sequences (Yuan *et al.* 1998; Turner *et al.* 1999) any analysis of just the dominant sequence will be limited. The literature varies in positions ascribed to promoter and enhancer regions (outlined in Table 1.2.), therefore wherever possible correlations have been made with respect to specific binding motifs that are agreed upon in the literature (again see Tables 1.1 and 1.2). This discussion is based therefore on the literature cited and comparisons made with respect to the database I have collected (Appendix B and disc appendix 'Data base alignment' file).

4.1.1 Transcriptional control element variations

"Even a single base change in a module of an enhancer promoter sequence can lead to a change in function" (Ogata *et al.* 1993). In light of this statement it is necessary to examine mutations that occur in my 16 samples.

S I promoter

The Ton-Is sequence contains a unique mutation in the TATA box which could decrease binding affinity, and thus promoter activity, thereby reducing HBs_L Ag production.

preS II promoter

NZM-7 and NZM-10 have unique mutations within this region. However, they are at sites that exhibit variability in the database implying a lack of constraints at those sites, and in the case of NZM-10 outside known functional sequence motifs, thus these probably have minimal affects. Sample NZM-2 however, has interesting dual mutations due to the overlapping reading frames. These affect the preS II/S promoter, an amino acid within the polymerase spacer region and the HBs_LAg. The nucleotide mutations affect two consecutive amino acid codons (bp position 3058, and 3062) which are not in known functional sequence motifs for the preS II/S promoter region but they do result in changes in the HBs_LAg coding sequence. An obvious assumption is therefore that they may affect the interplay between the virus and immune response.

Enhancer I and X promoter region

The boundaries to each region and relevant sequence motifs overlap, therefore these two transcriptional control elements will be considered together.

Samples Ni/To-N, NZM-Ma, Sam-En, NZM-2, NZM-3, NZM-5, NZM-9, Ton-Is, and NZM-8 all demonstrated variation within the regions. Most changes though were at sites that were highly variable within the database implying a lack of constraint at those particular sites. However, the NZM-8 mutations (bps 1116 and 1118) are unique and in conserved sites. The mutations correspond to an amino acid change in the overlapping region of the polymerase reverse transcriptase gene. These mutations could effect the polymerase function and also X gene production. As the X protein is a transacting activator of the genomic transcripts, any change in its production will in turn affect genomic transcript production. However, these mutations are not within known functional motifs therefore are unlikely to have any significant affect.

The unique mutation at base 1010 of NZM-9 is in a conserved site but outside any observed functional sequence motif. Ton-Is has two unique mutations at bases 970 and 977 that are both in conserved sites. In addition, bp 977 falls within a C/EBP binding region. The mutations therefore could potentially interfere with the interactions of this complex altering the expression of the core/pregenome and X protein.

Enhancer II region and Core/pregenomic promoter

Again the boundaries of each region and their relevant sequence motifs overlap; it is therefore simplest to consider these two transcriptional control elements together.

The enhancer II has been divided up into two transcriptional factor binding-sites, α and β. The α region has been broadly defined to lie between bases 1684-1733, but loss of sequence between 1706-1733 will result in complete loss of the α motif enhancer II activity (Lopez-Cabrera *et al.* 1991; Yuh and Ting 1991; Chen *et al.* 1993). Of the Pacific samples, Sam-En and Ton-Is have mutations in this region, Sam-En at base 1723 and Ton-Is at bases 1713 and 1716. While it is doubtful that these mutations prohibit binding, they may decrease efficiency resulting in a down regulation of the core/pregenomic transcript.

The β motif corresponds to a C/EBP (CCAAT/enhancer binding protein) binding site it has been demonstrated to have differential control between the enhancer II and the core/pregenomic promoter with high concentrations of the protein inhibiting the promoter activity (Yuh and Ting 1991; Chen *et al.* 1993;). The α motif is basal to expression and essential for enhancer activity, while the C/EBP motif is accessory and regulates expression.

All of the sequences have variations within the β region (bases 1744-1774). The most pronounced of these are samples NZM-1 and NZM-Ma, both of which have part of the binding motif missing due to an 8 bp deletion. This deletion mutation is not unique and is well reported in the literature (Horikita *et al.* 1994; Okamoto *et al.* 1994; Uchida *et al.* 1995). It has a two fold effect; truncating the X protein which is a trans-activating promoter and altering the balance between the core/pregenomic promoter. Lack of expression of HBeAg has been directly related to mutations in the β motif such as these deletions (Lopez-Cabrera *et al.* 1991; Moriyama *et al.* 1994; Uchida *et al.* 1995; Yuan *et al.* 1998) however, NZM-1 and NZM-8 were both HBeAg positive. Functional characterization of similar deletions has shown that they can be 'rescued in trans' by wild type strains or other members of the quasi-distribution (Yuan *et al.* 1998; Turner *et al.* 1999). So while this genome is defective for the production of HBeAg, the infected individual is not due to the quasi-species distribution within the host.

With regard to the other samples, only NZM-2 and NZM-8 demonstrate mutations in the C/EBP consensus binding motif, and are HBeAg positive. The NZM-2 mutation (bp 1750) has not been specifically reported before, but the NZM-8 has (Moriyama *et al.* 1994) and was associated with a HBeAg negative state. As the sequences obtained are the predominant sequence it is apparent in the NZM-8 case, and probably in the NZM-2 case, that these genomes are being 'rescued in trans 'by other genomes within the host such that the HBeAg state is maintained.

In the multiple footprint bindings sites associated with the core/pregenomic promoter region bases 1598-1881 (two of which are C/EBP sites (FP3 and FP6) (Lopez-Cabrera *et al.* 1990) the sequences demonstrate variations that are genotypically specific, therefore are not as a direct result of mutation.

NRE

No mutations within functional motif.

DR1 and DR2

No mutations

e

There is a single unique mutation in sample NZM-5 at base 1915, which could potentially affect encapsulation and synthesis of viral DNA (Asahina *et al.* 1996b). However it is outside the stem loop structure and not within predetermined binding motifs. NZM-3, NZM-4, NZM-5, NZM-8, To/En-V and To/n-Is all have the 1896 G to A mutation, and NZM-3 and NZM-10, have the 1899 G to A mutation. These two mutations create a more stable stem-loop structure that can enhance both viral replication and encapsulation (Uchida *et al.* 1995; Asahina *et al.* 1996b; Miyakawa *et al.* 1997).

Precore stop codon

All HBeAg negative samples have the 1896 precore stop codon. The G to A 1896 mutation creates a stop codon that stops the production of the HBeAg protein. Presence of this antigen correlates directly with HBV DNA titer in the blood stream, where HBeAg negative implies a comparatively lower titer than HBeAg positive (Chisari and Ferrari 1995). It is very interesting to find that Sam-En has the stop codon mutation but is HBeAg positive. It is possible that this is an artifact of the quasi-species distribution, however, this is the only case that I am aware of where the mutation is present and the individual is HBeAg positive.

4.1.2 Viral gene variations

X protein

The X protein trans-activates homologous and heterologous transcriptional regulatory elements including the enhancer I and core/pregenomic promoter. Therefore, the 8 bp deletions of NZM-1 and NZM-Ma (as previously mentioned) will result in the lack of expression of the HBV pregenome and genomic transcripts (Uchida *et al.* 1995). This mutation has an obvious affect which can be 'rescued in trans'. The other mutations resulting in amino acid changes in samples NZM-3 (aa X-136), NZM-7 (aa X-80), Sam-En (aa X-62, X-117) and Ton-V (aa X-102) do not have such an obvious result. They may affect the binding of the X protein to specific motifs, but as no studies on the binding interactions could be found, no conclusions can be drawn.

Polymerase

•*Terminal protein.*

The terminal protein primes reverse transcription and any amino acid changes may increase or decrease binding affinity (Asahina *et al.* 1996b). While the variations within this region are predominately genotype specific, unique variations were found in NZM-3 (aa P-171) and NZM-7 (aa P-49). A lack of function of the terminal protein will result in no reverse transcriptase priming, and no packaging of the genome.

•*Spacer.*

The samples have multiple amino acid changes in the spacer region, but as this region has no protein coding function, these should not affect any function or expression of the genome.

•*DNA polymerase/reverse transcriptase.*

Multiple amino acid changes, both unique and common, were identified. Once again, while lack of function of the polymerase enzyme has an obvious effect on elementary viral replication, as far as I am aware no mutational analysis of the HBV polymerase/reverse transcriptase has been reported. Therefore, no interpretation of the variations in the protein can be discussed, except perhaps to say these mutations may decease polymerase activity.

•Ribonuclease H

NZM-1 contains a stop codon truncating the protein by 65 amino acids at aa P-781. The RNaseH functions as a exonuclease, degrading viral mRNA template. A defunct RNAseH would result in lack of degradation of the pregenomic viral mRNA within the virion. This is unlikely to affect the infectious ability of the virion.

Core protein

The core protein serves as the envelope for the viral capsid. Mutations will effect the folding structure and deletions have been reported to result in the lack of expression of HBcAg (Yuan *et al.* 1998). NZM-Ma has a 33 base pair mutation that is within a well reported deletion range (amino acids 80 to 120 of the core protein) that results in the lack of expression of HBcAg. Therefore, this genomic sequence cannot produce the core protein, and on its own is replication defective. However, in the presence of other strains without the mutation it can replicate (Yuan *et al.* 1998). The core protein also has antigenic properties and as the deletion coincides with a HLA class I and class II - restricted T-cell epitope it may aid the virus to escape an immune response (Chisari and Ferrari. 1995; Yuan *et al.* 1998).

The other samples displayed mutations within the core protein (NZM-3, NZM-7, NZM-8, NZM-9, NZM-10), these mutations could quite possibly be part of a response to immunological factors which may be a factor contributing to the asymptomatic state of the infected individuals.

Surface Antigen.

There is a range of unique and common variations within the samples NZM-2, NZM-3, NZM-4 NZM-5, NZM-7, Sam-En, and To/En-V, affecting all the surface antigens. These could also possibly be part of a response to immunological factors which may be a element contributing to the asymptomatic state of the infected individuals. The methionie to isoleucine mutation at the start site of the preS II in NZM-3 would presumably affect the production of the $HBs_{M}Ag$ it is an interesting question that serological tests could answer.

Dual Genotype infection. NZM-8

Dual infection has been reported elsewhere (Bollyky *et al.* 1996) but is still an interesting phenomenon to find. How it occurred in this case is uncertain but with routine procedures such as blood transfusions it is not inconceivable.

4.1.3 Summary

These sequences demonstrate the classic inconsistency for which HBV is notorious (Karayiannis 1995). All donors were asymptomatic with varying levels of liver damage, and there was no significant correlation between their genomic mutations and clinical state. Mutations that occur in the more severe cases of liver disease also occur in those with no liver damage, for example, all samples had mutations in the β motif of enhancer II. In addition, samples NZM-1, NZM-3, NZM-7, NZM-Ma, Sam-En, and Ton-V all have mutations in the X protein but show different levels of liver damage. Further, when comparing mutations in these samples with those reported in the literature and the associated clinical states there is little consistency. The precore stop codon is the first obvious examples of this with the mutation commonly associated with fulminant and acute HBV infection, but 6 out of the 16 asymptomatic carriers had it. The 8bp deletion found in NZM-1 and NZM-Ma is another example; it should result in a negative HBeAg state (Uchida *et al.* 1995; Yuan *et al.* 1998), but both NZM-1 and NZM-8 were HBeAg positive. Moreover the deletion was found in association with acute and chronic HBV infection (Uchida *et al.* 1995) but NZM-Ma was not displaying any liver damage. As previously stated, the explanation is almost certainly due to an affect of the quasi-species distribution that is apparent in these two samples. Finally there were two wild-type sequences Sam-F and Ni/To-N displaying the same clinical state as two non wild-type sequences, To/Eu-V and Ton-Is. As all of the infected individuals were asymptomatic for viral infection and had unique variations in their viral sequences (two were wild-type), no correlation between sequences and diseases state could made from the data available.

4.2 Hepadnaviridae phylogenetic analysis

4.2.1 Pacific HBV samples

The samples were found unequivocally to be of genotype HBV-**C** and HBV-**D**. All criteria were in agreement and supported by strong bootstrap, and nearest neighbor bootstrap values.

The Pacific samples that were found to be of genotype HBV-**C**, group with the Pacific-specific strains from New Caledonia and French Polynesia (Norder *et al.* 1994) (see Table 3.5. and Figure 3.2), confirming them to be Pacific-specific strains. The Maori and the Tongan samples separate into two sub-group clusters suggesting that two viral sub-lineages have developed over time due to isolation. This leads to an interesting correlation between the HBV phylogenies and their Polynesian hosts. The viral strains present in Polynesia are a Pacific-specific sub-grouping of the Asian HBV-**C** strains. The Lapita people migrated into the Pacific from South-east Asia about 5000 ybp.

The implication from this is that these Pacific HBV-**C** strains originated in South-east Asia and were carried into Polynesia by the Lapita peoples during their migration and therefore are indigenous to the Polynesian people. The separation of the Tongan and Maori HBV-**C** samples supports this interpretation.

The Pacific HBV-**D** genotypes also demonstrated intriguing phylogenic characteristics. The general consensus is that the HBV-**D** genotype originated in the Mediterranean and has since become distributed world wide (Norder *et al.* 1992, 1993, 1994, 1996; Magnius and Norder 1995;). The strains I have sequenced form two distinct clusters, one shallow grouping, and another deeper divergence (Appendix C) within HBV-**D** genotypes. Because no rate of evolution has been determined for this virus (Bollyky and Holmes 1998), the time of this split is indeterminable. It is possible that this deeper branch within the HBV-D genotype represents a 300 year old import from the first European visitors to New Zealand and the South Pacific. Alternatively, as HBV-D has been found in South east Asia, the Lapita people could have brought this lineage with them 5000 ybp. Resolution of these hypotheses requires greater viral sampling and characterization of HBV-**D** genotypes within the Pacific, followed by further phylogenetic analysis. Such a study could be useful in estimating the evolutionary rate of the virus.

An intriguing observation is that there is no HBV-**A** genotypes present in our sample set. it was expected to be present due to the European influence in the Pacific. This sample set is not large enough to determine whether HBV-**A** is actually absent from the Pacific, or whether our sample set was to small to contain them. A wider study including Europeans might demonstrate HBV-**A** presence in the Pacific. This is something to consider in future studies.

4.2.2. Covarion variation analysis

The reason for investigating this mechanism of evolution is based upon the three hypotheses that could explain the difficulty in determining the position of the root in the HBV evolutionary tree. For example, genotype **F** may be genuinely older, have a faster rate, or have a change in covarion structure.

To determine whether HBV-**F** is genuinely older, greater taxa sampling of the genotype is required, however, there were only three complete genomes present on the Genbank database at time of analysis. More taxa may demonstrate a less extreme distance between HBV-**F** and the other genotypes shortening the long branch, or alternatively additional taxa may support the separation. Investigation of the different rates requires determination of an evolutionary rate for each genotype. Presently no rate for any has been determined.

Such analysis requires the long term study of asymptomatic HBeAg positive carriers through generations (Bollyky and Holmes 1998). Previous studies have mainly investigated the short-term mutational potential of the virus within a host, not the long-term mutations maintained within a population which is the actual mutation rate of the virus. The Whakatane child health and hepatitis foundation has a database of HBV blood samples (set up in 1984) which has the potential to provide the required data. They have screened an entire population to pick up asymptomatic carriers at an early date, and have followed these carriers through to the present day. Investigation of this hypothesis using these samples is recommended, and a project has been initiated for such a study.

Change in covarion structure can be investigated using the covarion integrity test. This test is simple to set up, requiring only an database alignment of the different genotypes. Considering the specific genotypic variations within the genomes change in covarion structure is a plausible reason why there is a lack of resolution in the HBV evolutionary tree. The analysis however, did not suggest a variation in covarion structure at all; indeed it suggested the opposite, that there is no variation in covarion structure between the genotypes.

4.2.3 Summary

The phylogenetic analysis of the Pacific samples has revealed interesting results. HBV genotypes HBV-**C** and HBV-**D** have been identified as present in the South Pacific, a previously undetermined fact. Moreover the evolutionary pattern of the virus also appeared to match the migratory pattern of their human host, a relationship which has not been considered before in HBV phylogenetic analysis. It is wonderfully curious to find that the HBV-**C** genotype demonstrates an evolutionary pattern that resembles its Pacific host. Determination of whether this is a true pattern is a future direction for our groups research. The results of the covarion inequality have also provided, in part, justification for research with the Whakatane child health foundation, a project which will among other things, investigate the evolutionary rate of HBV.

4.3 HLA-DPA1 allele typing data

The sequenced based typing technique was established as starting point for future genetic analysis of populations. The technique was shown to be simple, quick, definitive and with the ability to determine new variants. It is an encouraging result for future analysis. The HLA DPA1 data highlights the homogeneity of this population of the Trobriand Islanders. It has been established through concurrent analysis of different genetic population markers, that the Trobriand are of Asian descent and are highly homogenious (Nagy *et al.* 1996; Hagelberg *et al.* 1999a, 1999b).

The present data agrees with these studies, and is presented as a paper in Appendix D. Eighty six percent of the population is homogenous for the HLA DPA 02022, an allele which is common in Samoans and is a typical variant of Asians (Velickovic *et al.* 1998). The result with the Trobriand Islander samples is encouraging in the sense that it will probably be possible, as more data becomes available, to develop a quantitative model of human migration into Oceania. The Melanesian region has high genetic diversity, but our results are among those that demonstrate that there is high regional differentiation (Hagelberg *et al.* 1999a, 1999b). The alternative could have been that the diversity was randomly spread throughout Melanesia. But it is not.

The fact that the Trobriand Islanders are still genetically distinct, appears to support the fast train model for settlement of Polynesia (Bellwood 1987; Kirch 1997; Spriggs 1997). That is that the founding population that eventually settled Polynesia moved very quickly (in evolutionary terms) across Melanesia, with relatively little genetic admixture over that time period. However, there could have been more admixture later after Polynesia was already settled. Although the present results are reasonably consistent with the fast-train model, the model is probably an over simplification. In particular, the Fiji region is probably more complex (Ium *et al.* 1994), and there is virtually no sequence data available from there. Thus a priority for future work would be to obtain samples from this region.

4.4 Conclusions

This study began as an exploratory investigation of HBV in the Pacific, including the genetics of both the host and pathogen. An extensive literature review demonstrated the plight of the Polynesian people with this disease, it is hyper-endemic and the most common cause of cancer for males in Fiji (Mathi and Krishna 1998). Techniques and protocols for sequencing the genomes have been established within our lab for future research and have proved to be very informative. The sequences obtained contained a range of mutations with samples NZM-1 and NZM-Ma highlighting the potential affects of the quasi-species distribution within the individual. The presence of both HBV-**C** and HBV-**D** genotypes in the Pacific and the curious absence of HBV-**A** were established, these facts were previously unknown. It was also found through the covarion inequality test that there is not likely to be a change in covarion structure between the genotypes. A direct result of this study's findings has therefore been the establishment of research project into the evolutionary rate of HBV with the Whakatane child health, hepatitis foundation. The HLA allele typing proved to be an informative technique that is now an established technique with in our lab for future research. It has provided basic data for future studies on host pathogen interaction. As an exploratory study this masters project has provided a wealth of information and indicated future research directions.

4.5 Skills

The overall aim of this study was to investigate HBV and the HLA alleles in the Pacific; to establish protocols and directions for future research. From personal viewpoint the following skills have been established with in this project

- DNA extraction under biohazard containment of human and viral DNA
- Primer design using Oligo 4 software
- Establishment of PCR protocol in our lab for both HLA-DPA 1 and the complete HBV genome
- Extensive trouble shooting skills in the advent that the PCR template does not amplify.
- PCR product purification
- Cloning
- High accuracy DNA sequencing (allowing the identification of heterozygotes)
- Familiarity with currently available ABI sequencing analysis software (see table 2.8)
- Sequence alignment automated (Clustal X) and manual (Se-Al)
- Phylogenetic analysis of the data and usage of the application programs associated with the analysis (see table 2.8)
- Interpretation of the data and phylogenetic signals within it
- Word processing and graphic applications such as Mac Pro
- Design, interpretation, and thinking through experiments.

Thus, not only has this project provided interesting results but also has been very successful from a personal point of view.

Bibliography

Adams, D. The Hitch-hikers' guide to the Galaxy. Original radio series, 1978.

Alexopoulou, A., Karayiannis, P., Hadziyannis, S.J., Hou, J., Pickering, J., Luo, K., and Thomas, H.C. 1996. Whole genome analysis of hepatitis B virus from four cases of fulminant hepatitis: genetic variability and its potential role in disease pathogenicity. *Journal of Viral Hepatitis.* **3:** 173-181.

Alexopoulou, A., Karayiannis, P., Hadziyannis, S.J., Aiba, N., and Thomas, H.C. 1997. Emergence and selection of HBV variants in an anti-HBe positive patient persistently infected with quasi-species. *Journal of Hepatology.* **26:** 748-753.

Arauz-Ruiz, P. Norder, H., Visona, K.A. and Magnius, L.O., 1997. Molecular epidemiology of hepatitis B virus in central America reflected in the genetic variability of the small S gene. *The Journal of Infectious Diseases.* **176:** 851-858.

Asahina, Y., Enomoto, N., Ogura, Y., Kurosaki, M., Sakuma, I., Izumi, N., Marumo, F., and Sato, C. 1996a. Sequential changes in full-length genomes of hepatitis B Virus accompanying acute exacerbation of chronic Hepatitis B. *Journal of Hepatology.* **25**(6): 787-794.

Asahina, Y., Enomoto, N., Ogura, Y., Sakuma, I., Kurosaki, M., Izumi, N., Marumo, F., and Sato, C. 1996b. Complete nucleotide sequence of hepatitis B virus genomes associated with epidemic fulminant hepatitis. *Journal of Medical Virology.* **48**(2): 171-178.

Austin, C.C. 1999. Lizards took the express train to Polynesia. *Nature.* **397(**6715): 113

Bandelt, H-J. and Dress, A.W.M. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogeny and Evolution.* **1:** 242-252.

Bellwood, P. 1987. The Polynesians: Prehistory of an island people. Hampshire: Thames and Hudson.

Bellwood, P. 1991. The Austronesian dispersal and the origin of languages. *Scientific American.* **265**(1): 70-75.

Bichko, V., Pushko, P., Dreilina, D., Pumpen, P., and Gren, E. 1985. Subtype ayw variant of hepatitis B virus. DNA primary structure analysis. *FEBS Letters.* **185**(1): 208-212.

Bollyky, P.L. Rambaut, A., Harvey, P.H. and Holmes, E.C. 1996. Recombination between sequences of Hepatitis B virus from different genotypes. *Journal of Molecular Evolution.* **42:** 97-102.

Bollyky, P.L., and Holmes, E.C. 1998. Reconstructing evolutionary history of hepatitis B virus. *Journal of Molecular Evolution. (In pres)s.*

Bozkaya, H., Ayola, B., and Lok, A.S.F. 1996. High rate of mutations in the hepatitis B core gene during the immune clearance phase of chronic hepatitis B infection. *Hepatology.* **24:** 32-37.

Bozkaya, H., Akarca, U.S., Ayola, B., and Lok, A.S.F. 1997. High rate of conservation in the hepatitis B core gene during the immune clearance phase of chronic hepatitis B virus infection. *Journal of Hepatology.* **26:** 508-516.

Cann, R.L,. and Wilson, A.C. 1983. Length mutations in human mitochondrial DNA. *Genetics.* **104:** 699-711.

Chen, S-T., La Porte, P., and Yee J-K. 1993. Mutational analysis of Hepatitis B virus Enhancer 2. *Virology* **196:** 652-659.

Chisari, F.V. and Ferrari, C. 1995. Hepatitis B virus immunopathogenesis. *Annual Review of Immunology.* **13:** 29-60.

Cohen, J.I., Miller, R.H., Rosenblum, B., Denniston, K., Gerin, J.L. and Purcell, R.H. 1988. Sequence comparison of woodchuck hepatitis virus replicative forms shows conservation of the genome *Virology.* **162**: 12-20.

Courouce-Pauty, A.M., Plancon, A. and Soulier, J.P. 1983. Distribution of HBsAg subtypes in the world. *Vox-Sang* . **44**(4): 197-211.

Department of Survey and Land information, Wellington, New Zealand. 1994. Infomap 343 Nations of the world edition 6. Robertson's projections.

Deka, R., Jin, L., Shriver, M.D., Yu, L.M., Saha, N., Barrantes, R., Chakraborty, R., and Ferrell, R.E. 1996. Dispersion of human Y chromosome haplotypes - Based on five microsatellites in global populations. *Genome Research.* **6:** 1177-1184.

Dobson, A.P., and Carper, E.R. 1996. Infectious diseases and human population history. *BioScience.* **46:** 115-125.

Dikstein, R., Faktor, O., Ben-Levy, R., and Shaul, Y. 1990. Functional organization of the hepatitis B Virus enhancer. *Molecular and Cellular Biology.* **10**(7): 3683-3689.

Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans. CBMS-NSF regional conference series in applied mathematics, monograph no. 38. Society of industrial and applied mathematics, Philadelphia, PA.

Estacio, R.C., Chavez, C.C., Okamoto, H., Lingao, A.L., and Reyes, M.T. 1998. Nucleotide sequence of a hepatitis B virus genome subtype adw Isolated from a Philippino: Comparison with the reported three genomes of the same subtype. *Journal of Gastroenterology and Hepatology.* **3:** 215-222.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution.* **17:** 368-376

Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution.* **38**(1):16-24.

Felsenstein. J. 1993. *Phylip* (Phylogeny Inference package) Version 3.5. Computer software package and manual. Distributed by the author. Department of Genetics, University of Washington, Seattle, Washington.

Fujiyama, A., Miyanohara, A., Nozaki, C., Yoneyama, T., Ohtomo, N., and Matsubara, K. 1983 Cloning and structural analysis of Hepatitis B virus DNA, subtype *adr. Nucleic Acids Research.* **11**(13): 4601-4609.

Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. and Charnay, P. 1979. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. *Nature.* **281** (5733): 646-650.

Galibert, F., Chen, T.N. and Mandart, E. 1982. Nucleotide sequence of a cloned woodchuck hepatitis virus genome: Comparison with the hepatitis B virus sequence *Journal of Virology.* **41:** 51-65.

Gao, X., Zimmet, P., Serjeantson, S.W., 1992. HLA-DR, DQ Sequence polymorphisms in Polynesians, Micronesians, and Javanese. *Human Immunology.* **34:** 153-161.

Gibbons, A. 1994. Genes point to a new Identity for Pacific Pioneers. *Science.* **263:** 32-33.

Girones, R., Cote, P.J., Hornbuckle, W.E., Tennant, B.C., Gerin, J.L., Purcell, R.H. and Miller, R.H. 1989. Complete nucleotide sequence of a molecular clone of woodchuck hepatitis virus that is infectious in the natural host. *Proceedings of the National Academy of Science U.S.A.* **86:** 1846-1849

Gojobori, T., Moriyama, E.N., and Kimura, M. 1990. Molecular clock of viral evolution and the neutral theory. *Proceedings of the National Academy of Science USA.* **87:** 10015-10018.

Gunther, S. Meisel, H., Reip, A., Miska, S., Kruger, D.H., and Will, H.1992. Frequent and rapid emergence of mutated pre-C sequences in HBV from e-antigen positive carriers who seroconvert to anti-HBe during interferon treatment. *Virology.* **187:** 271-279.

Gust, I.D. 1996. Epidemiology of hepatitis B infection in the Western Pacific and South East Asia. *Gut.* **38**(suppl 2)**:** S18-S23.

Hagelburg, E., Goldman, N., Lio, P., Whelan, S., Schienfenhovel, W., Clegg, J.B., and Bowen, D.K. 1999a. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proceedings of the Royal Society of London.* **266**: 1-8.

Hagelberg, E., Kayser, M., Nagy, M., Roewer, L., Zimdahl, H., Krawczak, M., Lio, P., and Schiefenhovel, W. 1999b. Molecular genetic evidence for the human settlement of the pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. *Philosophical transactions of the Royal society of London series B* . **353:** 1-12.

Harford, N., Cabezon, T., Colau, B., Delisse, A.M., Rutgers, T., and De Wilde, M. 1987. Construction and characterisation of a *Saccharomcyes cerevisiae* strain (RIT 4376) expressing of HBsAg. *Post graduate Medical Journal.* **63** (Suppl 2)**:** 65-79.

Harihara, S., Hirai, M., Suutou, Y., Shimizu, K., and Omoto, K. 1992. Frequency of a 9bp deletion in the mitochondrial DNA among Asian Populations. *Human Biology.* **64**(2)**:** 161-166.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by molecular clock of mitochondrial DNA. *Journal of Molecular Evolution.* **21:** 160-174.

Hasegawa K. 1994. Enhanced replication of a Hepatitis B Virus mutant associated with an epidemic fulminant Hepatitis. *Journal of Virology.* **68**(3)**:** 1651-1659.

Hendy, M.D., and Penny, D. 1982. Branch and Bound algorithms to determine minimal evolutionary trees. *Mathmatical Biosciences.* **59:** 277-290.

Hertzberg, M., Mickleson K.N.P., Serjeantson, S.W., Prior, J.F., and Trent, R.J. 1989. An Asian 9 base pair deletion of Mitochondrial DNA is frequently found in Polynesians. *American Journal of Human Genetics.* **44:** 504-510.

Hill, A.V.S. and Serjeantson, S.W. (Eds). 1989. The colonisation of the Pacific: A genetic trail. Oxford: Clarendon Press.

Holmes, E.C., Nee, S., Rambaut, Garnett, G.R., Harvey, P.H., 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical . Transactions of the Royal Society of London.* B **349:** 33-40.

Horikita, M., Itoh, S., Yamamoto, K., Shibayama, T., Tsuda, F., and Okamoto, H. 1994. Differences in the entire nucleotide sequence between hepatitis B virus genomes from carriers positive for antibody to hepatitis B e antigen with and with out active disease. *Journal of Medical Virology.* **44**(1): 96-103.

Hurles, M.E., Irven, C., Taylor, P.G., dos Santos, F.R., Macaulay, V., Loughlin, J., Jobling, M.A., and Sykes, B.C.. 1998. European Y chromosomes in the Polynesian gene pool: a contrast to the structure revealed by mitochondria.. *American journal of human genetics.* **63**(6): **1790-1793.**

Huson, D.H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics.* **14**(1): 68-73.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules, pp 21-132. In Munro, H.N. *ed.* Mammalian protein metabolism. Academic Press. New York.

Kan, K.W., Dozy, A.M., Trecartin, R., and Todd, D., 1977. Identification of a nondeletion defect in α-thalassemia. *New England Journal of Medicine.* **297:** 457-474.

Karayiannis, P., Alexopoulou, A., Hadziyannis, S., Thursz, M., Watts, R., Seito, S., and Thomas, H.C. 1995. Fulminant hepatitis associated with hepatitis B virus antigen-negative infection: Importance of host factors. *Hepatology.* **22**(6): 1628-1634.

Kayser, M., de Knijff, P., Dieltjes, P., Krawczak, M., Nagy, M., Zerjalt, Pandya, A., Tyler-Smith, C, and Roewer, L. 1997. Applications of microsatellite-based Y chromosome haplotyping. *Electrophoresis.* **18(**9): 1602-7.

Kim Y. S. 1985. Nucleotide sequence of the core antigen-coding region of Hepatitis B virus (subtype adr-k). Korean Journal of Biochemistry. 18: 304-312.

Kim, K.T., Hyun, S.W., Kim, Y.S. and Rho, H.M. 1988. Complete nucleotide sequence of hepatitis B virus (subtype adr) Korean Journal of . Biochemistry. **21:** 319-331.

Kikuchi, M., Onda, H., Ono, I. and Sasada, R. 1984. Novel DNA its preparation and transformant the same *Patent*: JP 1984074985-A.

Kirch, P.V. 1997. The Lapita Peoples: Ancestors of the oceanic world. Oxford: Blackwell.

Klontz K.C. 1987. A program to provide hepatitis B Immunoprophylaxis to infants born to HBsAg-Positive Asian and Pacific Island women. *The Western Journal of Medicine.* **146**(2): 195-199.

Kniskern, P.J., Hagopian, A., Burke, P., Schultz, L.D., Montgomery, D.L., Hurni, W.M., Ip, C.Y., Schulman, C.A., Maigetter, R.Z., Wampler, D.E., Kubek, D., Sitrin, R.D., West, D.J., Ellis, R.W., and Miller, W.J. 1994. Characterization and evaluation of a recombinant hepatitis B vaccine expressed in yeast defective for N-linked hyperglycosylation. *Vaccine.* **12**(11): 1021-1025.

Kobayashi, M. and Koike, K. 1984. Complete nucleotide sequence of Hepatitis B virus DNA of subtype *adr* and its conserved gene organisation. *Gene.* **30:** 227-232

Kodama, K., Ogasawara, N., Yoshikawa, H., and Murakami, S. 1985. Nucleotide sequence of a cloned Wood chuck hepatitis virus genome: Evolutional relationships between hepadnaviruses. *Journal of Virology.* **56:** 978-986.

Kumar, S., Tamura, K., Nei. M., 1993. MEGA: Molecular evolutionary genetics analysis, version 1.01. The Pennsylvania State University, University Park, PA 16802.

Lai, M.E., Melis, A., Mazzoleni, A.P., Farci, P., and Balestrieri, A. 1991. Sequence analysis of hepatitis B virus genome of a new mutant of ayw subtype isolated in Sardina. *Nucleic Acids Research.* **19**(18): 5078

Lanford, R.E., Chavez, D., Brasky, K.M., Burns, R.B. III and Rico-Hesse, R. 1998. Isolation of a hepadnavirus from the woolly monkey, a new world primate. *Proceedings of the National Academy of Science U.S.A.* **95**(10): 5757-5761.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., and Howe C.J. 1998. A Covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Molecular Biology and Evolution.* **15**(9): 1183-1188.

Loncarevic, I.F., Zentgraf, H. and Schroder, C. 1990. Sequence of a replication competent hepatitis B virus genome with a preX open reading frame *Nucleic Acids Research.* **18**(16): 4940.

Lopez-Cabrera, M., Letovsky, J., Hu, K., and Siddiqui, A. 1990. Multiple liver-specific factors bind to the hepatitis B virus core/pregenomic promoter: Trans-activation and repression by CCAAT/enhancer binding protein. *Proceedings of the National Accademy of Science. USA.* **87:** 5069-5073.

Lopez-Cabrera, M., Letovsky, J., Hu, K., and Siddiqui. 1991. Transcriptional Factor C/EBP binds to and trans-activates the enhancer element II of the Hepatitis B virus. *Virology.* **183:** 825-829.

Lorenz, J.G., and Smith, D.G. 1994. Distribution of the 9-bp mitochondrial DNA region V deletion among north American Indians. *Human Biology* **66**(5): 777-788.

Lum, J.K., Rickards, O., Ching, C., and Cann, R. 1994. Polynesian mitochondrial DNA reveal three deep maternal lineages clusters. *Human Biology.* **66**(4): 567-800.

Pluciennicak, A. 1994. Molecular cloning and sequencing of two complete genomes of polish isolates of Human Hepatitis B Virus. Submitted (02-AUG-1994) Microbiology and Virology Center Polish Academy of Sciences, Laboratory of Genetic Engineering, ul. Dabrowskiego 251, Lodz, Poland, 93-231

Magnius, L.O., and Norder, H. 1995. Subtypes, genotypes, and molecular epidemiology of the hepatitis B virus as reflected by sequence variability of the S-Gene. *Intervirology.* **38:** 24-34.

Mairias, N., Bayes, M., and Tyler-Smith, C. 1994. Highly informative compound haplotypes for the human Y chromosome. *Human Molecular Genetics.* 1994. **3**(1): 115-123.

Mandart, E., Kay, A., and Galibert, F. 1984. Nucleotide sequence of a cloned Duck hepatitis B virus genome: Comparison with Wood Chuck and Human hepatitis B virus sequences. *Journal of Virology.* **49**(3): 782-792.

Mason W.S. and Seeger C *(Eds)*. 1991. *Current Topics in Microbiology and Immunology, Hepadnaviruses, Molecular biology and Pathogenesis* ***168.*** Philadelphia Springer-Verlag.

Mathai, E. and Krishna, M. 1998. Hepatitis infection in the Pacific. *Pacific Health Dialog.* **5:** 142-146.

Matisoo-Smith, E., Roberts, R.M., Irwin, G.J., Allen, J.S., Pennt, D., and Lambert, D.M. 1998. Patterns of prehistoric human mobility in Polynesia indicated by mtDNA from the Pacific rat. *Proceedings of the National Academy of Science.* **95:** 15145-15150.

Merriweather, D.A., Clark, A.G., Ballinger, S.W., Schurr, T.G., Soodyall, H., Jenkins, T., Sherry, S.T., Wallace, D.C., 1991. The structure of Human mitochondrial DNA variation. *Journal of Molecular Evolution.* **33**: 543-555.

Milne, A., Rodgers, E., and Hopkirk, N. 1995. Hepatitis B vaccination of babies in Melanesia. *The Lancet.* **346:** 318-320

Miller, R.H., and Robinson, W.S. 1986. Common evolutionary origin of hepatitis B virus and retroviruses. *Proceedings of the National Academy of Science. USA.* **83:** 2531-2535.

Miyakawa, Y., Okamoto, H., Mayumi, M. 1997. The molecular basis of Hepatitis B *e* antigen (HBeAg)-negative infections. *Journal of Viral Hepatitis.* **4:** 1-8.

Moriyama, K., Takada, T., Tsutsumi, T., Fukada, K., Ishibashi, H., Niho, Y., and Maeda, Y. 1994. Mutations in the transcriptional regulatory region of the precore and core/ pregenome of hepatitis B virus with defective HBeAg production. *Fukuoka Acta Med.* **85**(11): 314-332.

Mukaide, M., Kumazawa, T., Hoshi, A., Kawaguchi, R., and Hikiji, K. 1992. The complete nucleotide sequence of hepatitis B virus, subtype *adr* (SADR) and phylogenetic analysis. *Nucleic Acids Research.* **20**(20): 6105.

Muraskin, W. 1995. The formative years. In *The war against hepatitis B: A history of the international task force on hepatitis B virus.* Philadelphia, University of Pennsylvania Press.

Murray-McIntosh, R.P., Scrimshaw, B.J., Hatfield, P.J., Penny, D. 1998. Testing migration patterns and estimating founding population size in Polynesia by using human mitochondrial DNA sequences. Proceedings of the National Academy of Sciences USA. **95:** 9047-9052.

Nagy, M., Kurger, C., Petra, A., Kayser, M., 1996. Are the Trobrianders emigrants of South East Asia ? *Human Immunology* **47:** 1-2.

Naumann, H., Schaefer, S., Tachibana Yoshida, C.F., Coimbra Gaspar, A.M., Repp, R., and Gerlich, W.H. 1993. Identification of a new hepatitis B virus (HBV) genotype from Brazil that expresses HBV surface antigen subtype *adw4*. *Journal of General Virology*. **74:** 1627-1632.

Norder, H., Hammas, B., Lofdahl, S., Courouce, A., and Magnius, L.O. 1992. Comparison of the amino acid sequences of nine different serotypes of hepatitis B surface antigen and genomic classification of the corresponding hepatitis B virus strains. *Journal of General Virology*. **73:** 1201-1208.

Norder, H., Hammas, B., Lee, S., Bile, K., Courouce, A., Mushahwar, I.K., and Magnius, L.O. 1993. Genetic relatedness of hepatitis B Viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen. *Journal of General Virology*. **74:** 1341-1348.

Norder, H., Courouce, A., and Magnius, L.O. 1994. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology*. **198:** 489-503.

Norder, H., Ebert, J.W., Fields, H.A., Mushahwar, I.K., and Magnius, L.O. 1996. Complete sequencing of a Gibbon hepatitis B virus genome reveals a unique genotype distantly related to the Chimpanzee hepatitis B virus. *Virology*. **218:** 214-223.

Ogata, N., Miller, R.H., Ishak, K.G., Purcell R.H. 1993. The complete nucleotide sequence of a precore mutant hepatitis B virus implicated in fulminant hepatitis and its biological characterization in Chimpanzees. *Virology*. **1994:** 263-276.

Okamoto, H., Imai, M., Shimozaki, M., Hoshi, Y., Iizuka, H., and Gotanda, T. 1986. Nucleotide sequence of a cloned Hepatitis B Virus genome, subtype *ayr*: Comparison with genomes of the other three subtypes. *Journal of General Virology.* **67:** 2305-2314.

Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R.I., Imai, M., Miyakawa, Y. and Mayumi, M. 1988. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *Journal of General Virology.* **69**(Pt 10): 2575-2583.

Okamoto, H., Tsuda, F., Akahane, Y., Sugai, Y., Yoshiba, M., Moriyama, K., Tanaka, T., Miyakawa, Y., and Mayumi, M. 1994. Hepatitis B virus with mutations in the core Promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. *Journal of Virology.* **86:** 8102-8110.

Onda, H. and Igarashi, K. 1983. Novel DNA, its preparation and host transformed with the same. *Patent.* JP 1983194897-A.

Onda, H. and Igarashi, K. 1984. DNA and its use. *Patent.* JP 1984080615-A.

Onda, H., Inada, T. and Igarashi, K. 1987. Novel DNA and Polypeptide. *Patent.* JP 1987187495-A.

Ono, Y., Onda, H., Sasada, R., Igarashi, K., Sugino, Y., and Nishioka, K. 1983. The complete nucleotide sequences of the cloned hepatitis B virus DNA; Subtype adr and adw. *Nucleic Acids Research* **11**(6): 1747-1757

Orito, E., Mizokami, M., Ina, Y., Moriyama, E.N., Kameshima, N., Yamamoto, M., and Gojobori, T., 1989. Host-independent evolution and a genetic classification of the Hepadnavirus family base on nucleotide sequences. *Proceedings of the National Academy of Science USA.* **86:** 7059-7062.

Palumbi, S.R., Martin, A., Romano, S., McMillan, W.O., Stice, L., and Grabowski, G. 1991. The simple fools guide to PCR, Version 2. University of Hawaii Zoology Department, Honolulu.

Park, G.T., Yi, Y.W., Choi, C.Y., and Rho, H.M. 1997. A negative regulatory element and it's binding protein in the upstream of enhancer II of hepatitis B virus. *DNA and Cell Biology.* **16**(12): 1459-1465.

Penny, D., Hendy, M.D., and Steel M.A. 1992. Progress with methods for constructing evolutionary trees. *Trends in Ecology and Evolution.* **7**(3): 73-79.

Penny, D., Hendy, M.D., Lockhart, P.J., and Steel, M.A. 1996. Corrected parsimony, minimum evolution, and hadamard conjugations. *Systematic Biology.* **45**(4): 596-606.

Penny, D., M^{c} Cormish, B., and Charleston, M., and Hendy M.D. 1999. Modelling the covarion model of Molecular evolution by hidden Markov chains. *Submitted.*

Petre, J., Van Wijnendaele, F., De Neys, B., Conrath, F., Van Opstal, O., Hauser, P., Rutgers, T., Cabezon, T., Capiau, C., Harford, N., De Wilde, M., Stephene, J., Carr, S., Hemling, H., and Swadesh, J. 1987. Development of a hepatitis B vaccine from transformed yeast cells. *Postgraduate Medical Journal.* **63**(Suppl 2): 73-81.

Preisler-Adams, S., Schlayer, H.J., Peters, T., Korp, P., and Rasenack, J. 1993. Complete nucleotide sequence of a hepatitis B virus, subtype *adw2* and identification of three types of C open reading frames. *Nucleic Acids Research.* **21**(9): 2258.

Pugh, J.C., and Bassendine, M.F. 1990. Molecular biology of hepadnavirus replication. In *Viral hepatitis. British medical Bulletin, A series of expert reviews.* **462**: 329-354. (Ed) Zuckerman, A.J. London: Churchill Livingstone.

Pult, I., Chouard, T.; Wieland, S.; Klemenz, R.; Yaniv, M., and Blum H.E. 1997. A hepatitis B virus mutant with a new hepatocyte nuclear factor 1 binding site emerging in transplant-transmitted fulminant hepatitis B.. *Hepatology* **25**(6): 1507-1515.

Redd, A.J., Takezaki, N., Sherry, S.T., M^{c}Garvey, S.T., Sofo, A.S., and Stoneking, M. 1995. Evolutionary history of the COII/tRNa^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Molecular Biology of Evolution* . **12**(4):604-615.

Renbao, G. Meijin, C., Lueping, S., Suwen, Q., and Zaiping, L. 1987. The complete nucleotide sequence of the cloned DNA of Hepatitis B virus subtype *adr* in pADR-1. *Scientia Sinica*. **30**(5): 507-521.

Riley E. and Olerup O. 1992. HLA Polymorphisms and Evolution. *Immunology Today*. **13**(9): 333-335

Roberts, R.G., Jones, R., and Smith, M.A. 1990. Thermoluminescence dating of a 50 000 year old human occupation site in Northern Australia. *Nature*. **345** 153-156.

Rho, H.M. Kim, K., Hyun, S.W., and Kim, Y.S. 1989. The nucleotide sequence and reading frame of a mutant Hepatitis B Virus subtype *adr*. *Nucleic Acids Research*. **17**(5): 2124.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Schare, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*. **239**: 487-491.

Saitou, N. and Nei, M. 1987. The Neighbor joining methods: A new method for reconstruction phylogenetic trees. *Molecular Biology and Evolution*. **4**(4) : 406-425.

Sastrosoewignjo, R.I., Omi, S., Okamoto, H., Mayumi, M., Rustam, M., and Sujudi, X. 1987. The complete nucleotide sequence of HBV DNA clone of subtype *adw* (pMND 122) from Mendo in Sulawesi Island, Indonesia. *ICMR Ann*. **7**: 51-60.

Saunders, P. (Eds). 1998. New Ethicals **35**:1 . Adif International. Auckland.

Serjeantson, S.W. 1989. HLA and Disease in Oceania. *Papau New Guinea Medical Journal*. **32**: 241-249.

Shiosaki, C., Takata, K.., Nishimura, S., Mizokami, H., and Matsubara, K. 1991. Production of hepatitis B virion-like particles in yeast. *Gene*. **106**: 143-149.

Soodyall, H., Vigilent, L., Hill, A.V., Stoneking, M., and Jenkins, T. 1996. Origins of the 9bp Deletion in Africans. *American Journal of Human Genetics*. **58**:595-608.

Spriggs, M. 1997. The Island Melanesians. Oxford: Blackwell.

Smith D.B. McAllister, J., Casino, C., and Simmonds, P. 1997. Virus "quasispecies": making a mountain out of a molehill ? *Journal of General Virology.* **78:** 1511-1519.

Stevens, G. 1985. Lands in collision, discovering New Zealand's past geography. Wellington. Science Information Publishing Centre.

Stoll-Becker, S., Repp, R., Glebe, D., Schaefer, S., Kreuder, J., Kann, M., Lampert, F., and Gerlich, W. H. 1997. Transcription of hepatitis B virus in peripheral blood mononuclear cells from persistently infected patients. *Journal of Virology.* **71:**5399-5407.

Studier, J.A. and Keppler, K.J. 1988. A note on the Neighbor Joining algorithm of Saitou and Nei. *Molecular Biology and Evolution.* **6**(5): 729-731.

Sung J.L. 1990. Hepatitis B virus eradication strategy for Aisa. *Vaccine.* **8**(supp): S95-S99.

Swafford, D.L. 1993. PAUP. Phylogenetic analysis using parsimony. Illinois natural history survey Campaign.

Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. 1996. Phylogenetic Inference. In Hillis, D.M., Moritz, C., and Mable, B.K. (Eds) Molecular Systematics. Second edition. Sinauer Associates, Inc., Sunderland, Massachusetts.

Swofford,D.L., 1998. PAUP*: Phylogenetic analysis using parsimony version 4.0.0d64. Illinois Natural History Survey Illinois.

Sykes, B., Leiboff, A., Low-Beer, J., Tetner, S., and Richards, M. 1995. The origins of the Polynesians: An interpretation from mitochondrial lineage analysis. *American Journal of Human Genetics.* **57:** 1463-1475.

Takemura, F., Ishii, T.; Fujii, N., and Uchida, T., 1990. Complete nucleotide sequence of hepatitis B virus. *Nucleic Acids Research.* **18**(15): 4587.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research.* **22:** 4673-80.

Tong, S., Li, J., Vitvitski, I., and Trepo, C. 1990. Active hepatitis B virus replication in the presence of anti-HBe is associated with viral variants containing an inactive pre-C region. *Virology.* **176:** 596-603.

Turner, P.E. and Chao, L. 1999. Prisoner's dilemma in an RNA virus. *Nature.* **398:** 441-443.

Uiriamu, J.R. and Hawaado, M.G 1981. Nontransit virus. *Patent*: JP 1981063995-A.

Uchida, T., Gotoh, K., and Shikata, T. 1995. Complete Nucleotide sequences and the characteristics of two hepatitis B virus mutants causing serologically negative acute or chronic Hepatitis B. *Journal of Medical Virology.* **45**(3): 247-252.

Vanenzuela, P., Medina, A., Rutter, W.J., Ammerer, G., and Hall, B.D. 1982. Synthesis and assembly of hepatitis B virus surface antigen particles in yeast. *Nature.* **298:** 347-350.

Vaudin, M., Wolstenholme, A.J., Tsiquaye, K.N., Zuckerman, A.J. and Harrison, T.J. 1988. The complete nucleotide sequence of the genome of a Hepatitis B Virus isolated from a naturally infected chimpanzee. *Journal of General Virology.* **69:** 1383-1389.

Velenzuela, P., Quiroga, M., Zalvidar, J., Gray, P., and Rutter, W.J. 1980. The nucleotide sequence of the hepatitis B viral genome and the identification of the major viral genes. *In Animal Virus Genetics.* 57-70 (Eds) Fields, B.N., Jaenisch, R., and Fox, C.F., New York. Academic Press.

Velickovic, Z.M., Harrison, G.L., Carter, J.M., Murray-McIntosh, R.P. 1998. HLA-DPA1 Sequencing based typing. *(In press).*

Versluis L.F. Rozemuller, E.H., Tonks, S., Marsh, S.G.E., Bouwens A.G.M., Bodmer, J.G., Tilanus, M.G.J. 1993. High resolution HLA-DPB typing based upon computerised analysis of data obtained by Fluorescent sequencing of the amplified polymorphic exon 2. Human Immunology. **38:** 227-283

Wallace, D.C. 1995. Mitochondrial DNA variation in human evolution, degenerative disease and aging. *American Journal of Human Genetics.* **57**: 201-223.

Wands,J.R., Liang,T.J., Blum,H.E. and Shafritz,D.A. 1992. Molecular pathogenesis of liver disease during persistent hepatitis B virus infection. Seminars in Liver Disease **12** (3): 252-264

Yuan, T.T-T., Lin M-H.,Qiu S.M. and Shih, C. 1998. Functional Characterization of naturally occurring variants of human hepatitis B virus containing the core internal deletion mutation. *Journal of Virology.* **72**(3): 2168-2176.

Yang Z., Lauder, I.J., and Lin, H.J. 1995. Molecular evolution of the hepatitis B virus genome. *Journal of Molecular Evolution.* **41**: 587-596.

Yuh, C-H., and Ting, L-P., 1991. C/EBP proteins binding to the functional box-α and box-β of the second enhancer of Hepatitis B virus. *Molecular and Cellular Biology.* **11**(10): 5044-5052.

Appendix A

Amino acid sequences for the proteins of the Pacific hepatitis B virus samples

Table A.1 HBeAg and HBcAg amino acid sequences

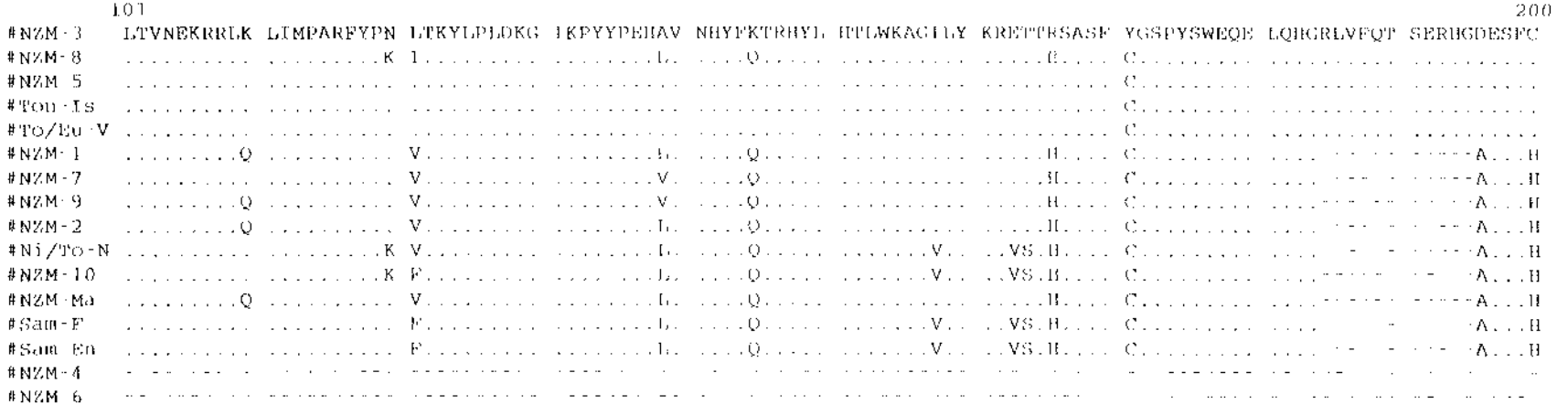
	1									100
#NZM-3	MQLFHLCLII	SCSCPTVQAS	KLCLGWL*DM	DIDPYKEFGA	SVELLSFLPS	DFFPSIRDLL	DTASALYREA	LESPEHCSPH	HTALRQAVLC	WGELMNLATW
#NZM-8	....	....	.......WG.	....	....	....	....	....E.F...	....	....
#NZM-5	....	....	........G.	...H......	....	....	....	....E.....	....	....
#Ton-Is	....	....	.......WG.	....	....	....	....	....E.....	....	....
#To/Eu-V	....	....	.......WG.	....	....	....	....	....E.....	....	....
#NZM-1	....	....	.......WG.	....	T.........	....V.....	....	....E.....	.......I..	.....T....
#NZM-7	....	....	........G.	....	....	....V.....	........D.	....E.....	.......I..	.....T....
#NZM-9	....	....	........G.	....	T.........	....V.....	....	....E...A.	.......I..	.....T....
#NZM-2	....	....	.......WG.	....	T.........	....V.....	....	....E.....	.......I..	.....T....
#Ni/To-N	....	....	.......WG.	....	T.........	....V.....	...A....D.	....E.....	.......I..	.....T....
#NZM-10	....	....	....	....	T.........	....V.....	...A....D.	....E...A.	.......I..	..D.......
#NZM-Ma	....	....	.......WG.	....	T.........	....V.....	....	....E.....	.......I..	.....T....
#Sam-F	....	....	.......WG.	....	T.........	....V.....	...A....D.	....E.....	.......I..	.....T....
#Sam-En	....	....	........G.	....	T.........	....V.....	...A....D.	....E.....	.......I..	..D..T....
#NZM-4	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------
#NZM-6	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

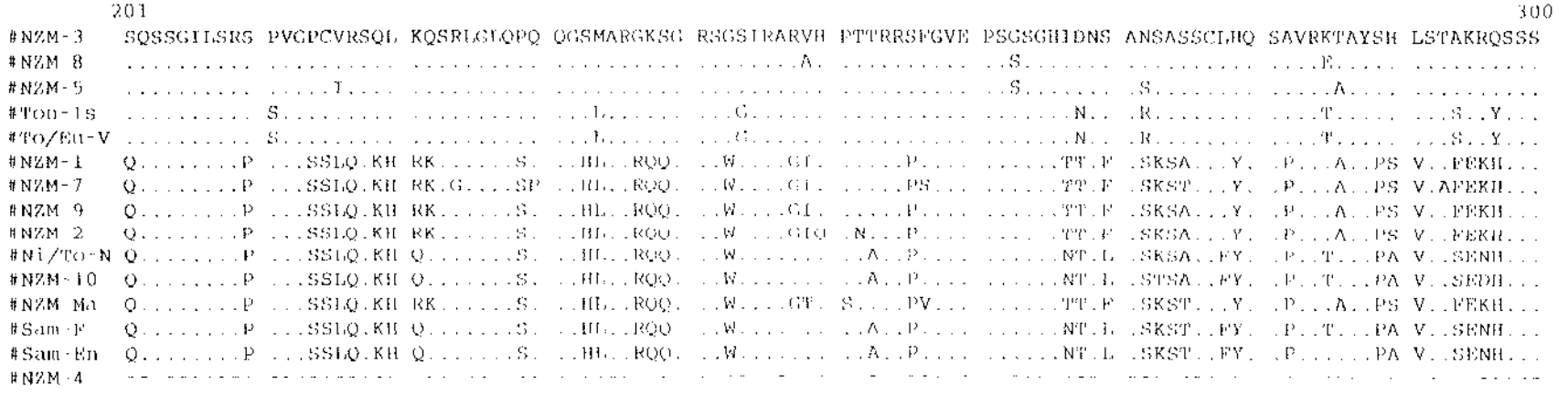
	101									200
#NZM-3	VGSNLEDPAS	RELVVSYVNI	NMGLKIRQLL	WFHISCLTFG	RETVLEYLVS	FGVWIRTPPA	YRPPNAPILS	TLPETTVVRR	-RGRSPRRR	TPSPRRRRSQ
#NZM-8	.......L..	....	.....I....	....	....	........P.	....	.......I..	....	....
#NZM-5	....	.....G....	.....I....	....	....	........P.	....	....	....	....
#Ton-Is	....	....	.....I....	....	....	........P.	....	....	....	....
#To/Eu-V	....	....	.....I....	....	....	........P.	....	....	....	....
#NZM-1	..G....I..	.D.......T	.....F....	....	....I.....	........P.	....	....	....	....
#NZM-7	..G....T..	.D.......T	.....F....	....	....V.....	........P.	...S......	.....C.I..	....	....
#NZM-9	..G....T..	.D.......T	.....F....	....	....	........P.	....	.......I.Q	.....T....	....
#NZM-2	..G....I..	.D.......T	.....F....	....	....I.....	........P.	....	....	....	....
#Ni/To-N	..T.......	.D.......T	.....F....	....	....	........P.	....	....	....	....
#NZM-10	..T....QP.	.D...G...T	TV...F....	....	....	........P.	....	.......I..	....	....
#NZM-Ma	..G....I..	.D........	.....F.H..	....	.G..I.....	........P.	....	....	....	....
#Sam-F	..T.......	.D.......T	.....F....	....	....	........P.	....	....	....	....
#Sam-En	..T.......	.D.......T	.V...F....	....	....	........P.	....	....	....	....
#NZM-4	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------
#NZM-6	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

```
            201            215
#NZM-3      SPRRRRSQSR ESQC*
#NZM 8      ........P. ..*.Q
#NZM 5      .......... .....
#Ton-1s     .......... .....
#To/Eu-V    .......... .....
#NZM-1      .......... .....
#NZM 7      .......... ..H..
#NZM-9      .......... .....
#NZM 2      .......... .....
#Ni/To-N    .......... .....
#NZM-10     .......... .P...
#NZM-Ma     .......... .....
#Sam-F      .......... .....
#Sam-En     .......... .....
#NZM 4      ---------- -----
#NZM 6      ---------- -----
```

Table A.2 Polymerase amino acid sequences

```
            1                                                                                                        100
#NZM-3      MPLSYQHFRK LLLLDD--EA GPLEEELPRL ADEGLNHRVA EDLNLGNLNV NIPWTHKVGN FTGLYSSTVP VFNPEWQTPS FPDIHLKEDI INRCQQYVGP
#NZM-8      .........R .......... .......... .......... .......... S......... .......... .......... .......... ..........
#NZM-5      .......... .......... .......... .......... .......... S......... .......... .......... .......... ..........
#Ton-1s     .......... .......... .......... .......... .......... S......... .......... .......... .......... ..........
#To/Eu-V    .......... .......... .......... .......... .......... S......... .......... .......... .......... ..........
#NZM-1      .........R .......... .......... ......R... .......... S......... .......... I...H.K... ..N..HQ... .KK.E.F...
#NZM-7      .........R .V........ .......... ......R... .......PT. .......... .......... I...H.K... ..N..HQ... .KK.E.F...
#NZM-9      .........R .....N.... .......... ......R... .......... .......... .......... I...H.K... ..N..HQ... .KK.E.F...
#NZM 2      .........R .......... .......... ......R... .......... S......... .......... I...H.K... ..N..HQ... .KK.E.F...
#Ni/To-N    .........R .......... .......... ...D..R... .......... S......... ......A... ....H.K... ..N..HQ... .KK.E.F...
#NZM-10     .........R .......... .......... ...D..R... .......P.. S......... .......... ....H.K... ..N..HQ... .KK.E.F...
#NZM-Ma     .........R .......... .......... ......R... .......... S......... .......... I...H.K... ..N..HQ... .KK.E.F...
#Sam-F      .........R .......... .......... ...D..R... .......... S......... ......A... ....H.K... ..N..HQ... .KK.E.F...
#Sam-En     .........R .......... .......... ...D..R... .......... S......... ......A... ....H.K... ..N..HQ... .KK.E.F...
#NZM-4      ---------- ------..-- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
#NZM 6      ---------- ------..-- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
```


101
200
#NZM-3 LTVNEKRRLK LIMPARFYPN LTKYLPLDKG IKPYYPEHAV NHYFKTRHYL IPLWKAGILY KRETTRSASF YGSPYSWEQE LQHGRLVFQT SERHGDESPC
#NZM-8
#NZM 5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM-9
#NZM-2
#Ni/To-N
#NZM-10
#NZM-Ma
#Sam-F
#Sam-En
#NZM-4
#NZM 6


201
300
#NZM-3 SQSSGILSRS PVGPCVRSQL KQSRLGLQPQ QGSMARGKSG RSGSIRARVH PTTRRSFGVE PSGSGHIDNS ANSASSCLHQ SAVRKTAYSH LSTAKRQSSS
#NZM 8
#NZM-5
#Ton-Is
#To/Eu-V
#NZM-1
#NZM-7
#NZM 9
#NZM 2
#Ni/To-N
#NZM-10
#NZM Ma
#Sam-F
#Sam-En
#NZM-4
#NZM-6

```
          301                                                                                                                  400
#NZM-3    GHTVNSTTFH QALLDPRVRG LYLPAGGSSS GTVNPVPTTA SPISSIFSRT GDPAPNNGEH NIRIPRTPAR VTGGVFLVDK NPHNTTESRL VVDFSQFSRG
#NZM-8    ..A....... .......... ..F....... .......... .......... .......... .......... .......... .......... ..........
#NZM-5    ..A....A.. .......... ..F....... .......... .....T.... .......... .......... .......... .......... ..........
#Ton-Ts   ..A....... .......... ..Y....... .......... .......... .......... .....G.... .......... .......... ..........
#To/Eu-V  ..AM...... .......... ..Y....... .......... .......... .......... .....S.... .......... .......... ..........
#NZM-1    ..A..... .T.Q...... ..F..... .....V .H.....I ....L..... H..... I..... .....A.... ..........
#NZM-7    ..A.....R .T.Q....K. ..F..... .....V .H.....I ....L..... H..... .....A.... ..........
#NZM-9    ..A..... .T.Q...... ..F..... .....V .H.....I ....L..... H..... .....A.... ..........
#NZM-2    ..A..... .T.Q...... ..F..... .....V .H.....I ....L..... H..... .....A.... ..........
#Ni/To-N  ..A..... .T.Q...... ..F..... .....V .H.....I ....L..... H..... .....A.... ..........
#NZM-10   ..A..... .T.Q...... ..F..... .....I ....L..... H..... .....A.... ..........
#NZM-Ma   ..A..... .T.Q...... ..F..... .....I ....L..... H..... .....A.... ..........
#Sam-F    ..A..... .T.Q...... ..F..... .....I ....L..... H..... .....A.... ..........
#Sam-En   ..A..... .T.Q...... ..F..... .....I ....L..... H..... .....A.... ..........
#NZM-4    ----..... .......... ..P..... .......... .......... .......... .......... .......... .......... ..........
#NZM-6    --  ..... .T.Q...... ..F..... .....V .H.....I ....L..... H..... .....A.... ..........

          401                                                                                                                  500
#NZM-3    STKVSWPKFA VPNLQSLTNL LSSNLSWLSL DVSAAFYHLP LHPAAMPHLL VGSSGLPRYV ARLSSTSRNI NYQHGTMQDL HDSGSRNLYV SLMLLYKTFG
#NZM-8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#NZM-5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Ts   ..R....... .......... .......... .......... .......... I......... .........L ..K....... .......... ..........
#To/Eu-V  ..R....... .......... .......... .......... .......... I......... .........L ..K....... .......... ..........
#NZM-1    NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
#NZM-7    NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
#NZM-9    NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
#NZM-2    NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
#Ni/To-N  NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
#NZM-10   NYR....... .......... .......... .......... .......... .....S.... .....N..I. DHK...L.N. .......... ..L.......
#NZM-Ma   NYR....... .......... .......... .......... .......... .....S.... .....N..I. .H....L.N. .......... ..L.......
#Sam-F    NYR....... .......... .......... .......... .......... .....S.... .....N..I. .H....L.N. .......... ..L.......
#Sam-En   NYR....... .......... .......... .......... .......... .....S.... .....N..I. .H....L.N. .......... ..L.......
#NZM-4    .......... .......... .......... .......... .....V.... .......... .......... .......... .......... ..........
#NZM-6    NYR....... .......... .......... .......... .......... .....S.... .....N..IP .H.....N. .......... ..L...Q...
```

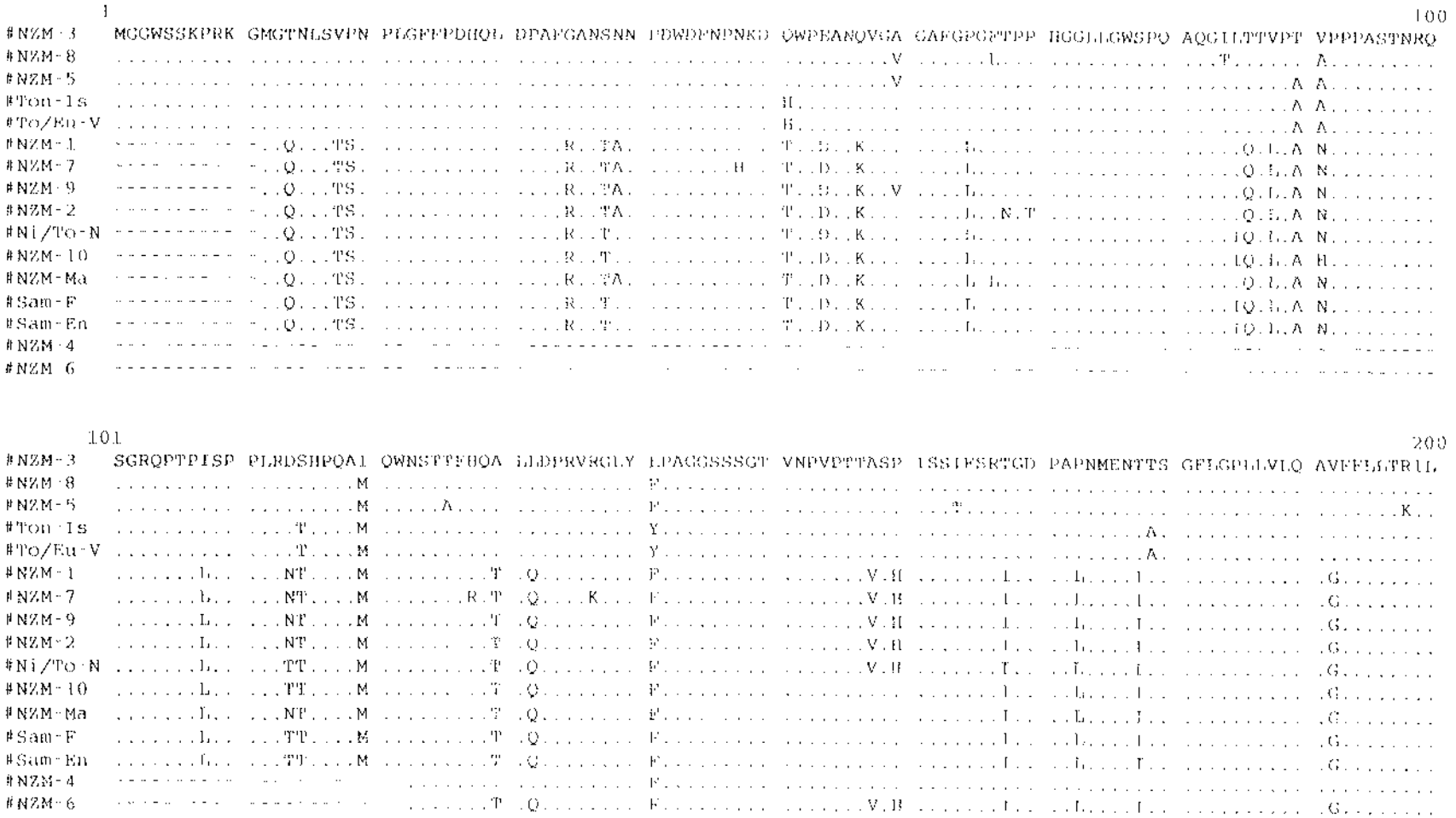
```
          501                                                                                                      600
#NZM-3    RKLHLYSHPI ILGFRKIPMG VGLSPFLLAQ FTSAICSVVR RAFPHCLAFS YMDDVVLGAK SVQHLESLYT SITNFLLSLG IHLNPNKTKR WGYSLNFMGY
#NZM-8    .......... .......... .......... .......... .......... .......... .......... AV........ .....Q.... ..........
#NZM-5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Is   .......... .......... .......... .......... .......... .......... .......... A......... .......... ..........
#To/Eu-V  .......... .......... .......... .......... .......... .......... .......... A......... .......... ..........
#NZM-1    .......... .......... .......... .......... .......... .......... ........F. AV........ .......... .....H....
#NZM-7    .......... .......... .......... .......... .......V.. .......... ........F. AV........ .......... .....H....
#NZM-9    .......... .......... .......... .......... .......... .......... ........F. AV........ .......... .....H....
#NZM-2    .......... .......... .......... .......... .......... .......... ........F. AV........ .......... .....H....
#Ni/To-N  .......... .......... .......... .......... .......... .......... ........F. AV........ .......... ..........
#NZM-10   .......... .......... .......... .......... .......... .........T ........F. AV........ .......... ..........
#NZM-Ma   .......... .......... .......... .......... .......... .......... ........F. AV........ .......... ..........
#Sam-F    .......... .......... .......... .......... .......... .......... ........F. AV........ .......... ..........
#Sam-En   .......... .......... .......... .......... .......... .......... ........F. AV........ .......... ..........
#NZM-4    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#NZM-6    .......... .......... .......... .......... .......... .......... ........F. AV........ .......... .....H....

          601                                                                                                      700
#NZM-3    VIGSWGTLPQ EHIFLKIKNC FRKLPVNRPI DWKVCQRIVG LLGFAAPFTQ CGYPALMPLY ACIQAKQAFT FSSTYKVFLC KQYLNLYPVA RQRSGLCQVF
#NZM-8    .......... ...V....H. .......... .......... .......... .......... .......... ......M... .......... ..........
#NZM-5    .......... ........H. .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Is   .......... ...VR...Q. .......K.. .......... .......... .......... .......... .......... .......... ..........
#To/Eu-V  .......... ...VR...Q. .......... .......... .......... .......... .......... .......... Q......... ..........
#NZM-1    ...CY.S... D...Q...E. .......... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-7    ...GF.S... D...Q...E. .......... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-9    ...CY.S... D...Q...E. .......... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-2    ...CY.S... D...Q...E. .......... .......... .......... .......... ....S..... ..P...A... ....T..... ...P......
#Ni/To-N  ...CY.S... D...Q...E. ......H... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-10   ...C..S... D...H...E. .......... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-Ma   ...C..S... D...H...E. ......H... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#Sam-F    ...C..S... N...H...E. ......H... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#Sam-En   ...C..S... N...H...E. ......H... .......... .......... .......... ....S..... ..P...A... .......... ...P......
#NZM-4    .......... ........H. .......... ---------- ---------- ---------- ---------- ---------- ---------- ----------
#NZM-6    ...CY.S--- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
```

```
           701                                                                                                   800
#NZM-3     ADATPTGWGL AICHQRMRGT FVAPLPIHTA ELLAACFARS RSGAKLIGTD NSVVLSRKYT SFPWLLGCAA NWILRGTSFV YVPSALNPAD DPSRGRLGIY
#NZM-8     .......... .......... .......... .......... .......... .......... .......... .......... .......... ........L.
#NZM-5     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Is    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#To/Eu-V   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#NZM-1     .......... VM........ .L........ .......... ....NIL... .......... .......... .......... *L........ ........LS
#NZM-7     .......... VM........ .P........ .......... ....NIL... .......... .......... .......... .......... ........LS
#NZM-9     .......... VM........ .L........ .......... ....NIL... .......... .......... .......... .......... ........LS
#NZM-2     .......... VM........ .L........ .......... ....NIL... .......... .......... .......... .......... ........LS
#Nj/To-N   .......... VM........ .L........ .......... ....HIL... .......... .......... .......... .......... ........LC
#NZM-10    .......... VM........ .L........ .......... ....NI.... .......... .......... .......... .......... ........P.
#NZM-Ma    .......... VM........ .L........ .......... ....HIL... .......... .......... .......... .......... ........PC
#Sam-F     .......... VM........ .L........ .......... ....NIL... .......... .......... .......... .......... ........PC
#Sam-En    .......... VM........ .L........ .......... ....NIL... .......... .......... .......... .......... ........PC
#NZM-4     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------
#NZM-6     ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

           801                                            846
#NZM-3     RPLLHLPFRP TTGRTSLYAV SPSVPSHLPD RVHFASPLHV AWRPP*
#NZM-8     ....R..... .........D .......... .......... ..K...
#NZM-5     ....R..... .......... .........G .......... ......
#Ton-Is    ....R..... .......... .........G .......... ......
#To/Eu-V   ....R..... .......... .........G .......... ......
#NZM-1     ....R..... .........D .......... .......... ......
#NZM-7     ....R..... .........D .......... .......... ......
#NZM-9     ....R..... .........D .......... .......... ......
#NZM-2     ....R..... .........D .......... .......... ......
#Nj/To-N   ....C..... .........D .........V .......... ......
#NZM-10    ....C..... .........D ......R..V .......... ......
#NZM-Ma    ....C..... .........D .........V .......... ......
#Sam-F     ....C..... .........D .......... .......... ......
#Sam-En    ....C..... .........D .......... .......... ......
#NZM-4     ---------- ---------- ---------- ---------- ------
#NZM-6     ---------- ---------- ---------- ---------- ------
```

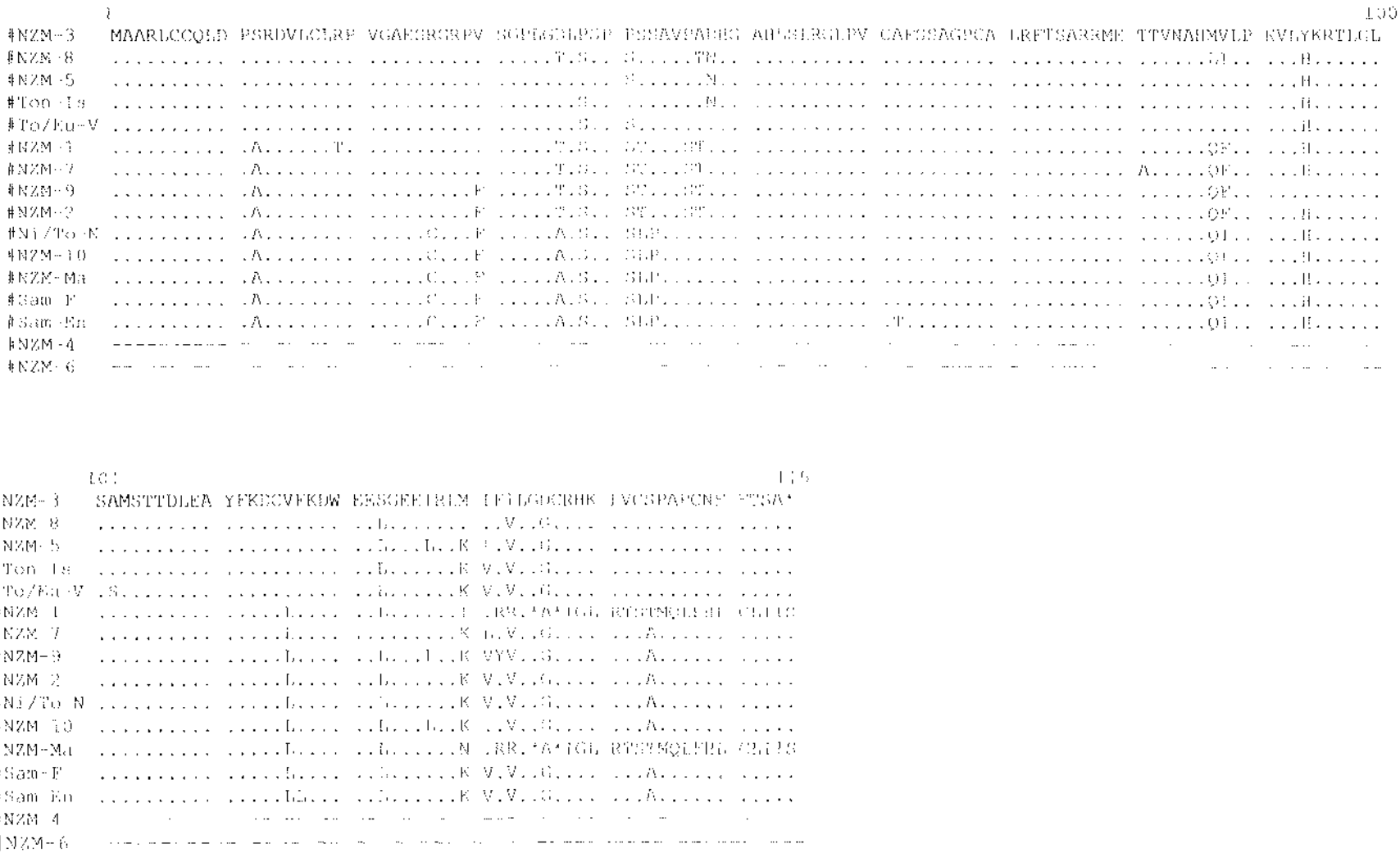
Table A.3 HBsAg amino acid sequences



```
           201                                                                                                     300
#NZM-3     TIPQSLDSWW TSLNFLGEAP RCPGQNSQSP TSNHSPTSCP PICPGYRWMC LRRFIIFLFI LLLCLIFLLV LLDYQGMLPV CPLLPGTSTT STGPCKTCTI
#NZM-8     .......... .......... .......... .......... .......... .......... .......... .......... .......... .........T
#NZM-5     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Ts    .......... .......... V......... .......... .T........ .......... .......... .......... .......... ..........
#To/Eu-V   .......... .......... V......... .......... .T........ .......... .......... .......... .......... ..........
#NZM-1     .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#NZM-7     .......... .......GTT V.R....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#NZM-9     .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#NZM-2     .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#Ni/To-N   .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#NZM-10    .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#NZM-Ma    .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#Sam-F     .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T
#Sam-En    .......... .......GTT V.L....... .......... .T.......Y .......... .......... .......... ...I..S... .....R...T
#NZM-4     .......... ...S...... K......... .......... .......... ..H....... .......... .......... .......... ..........
#NZM-6     .......... .......GTT V.L....... .......... .T........ .......... .......... .......... ...I..S... .....R...T

           301                                                                                                     401
#NZM-3     PAQGTSMFPS CCCTKPSDGN CTCIPIPSSW AFVRFLWEWA SVRFSWLSLL APFVQWFVGL SPTVWLSVIW MMWYWGPSLY NILSPFIPLL PIFFCLWVYI*
#NZM-8     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#NZM-5     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
#Ton-Ts    .......... .......... .......... .......... .......... .......A.. .......... .......... .......... ..........
#To/Eu-V   .......... .......... .......... .......... .......... .......A.. .......... .......... .......... ..........
#NZM-1     .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#NZM-7     .......Y.. .......... .......... ..GK...... .A........ V......... ......L... T......... ST....L... ..........
#NZM-9     .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#NZM-2     .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#Ni/To-N   .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#NZM-10    .......Y.. .......... .......... ..GK...... .A........ V......... .......... ........R. S.....L... ..........
#NZM-Ma    ..R....Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#Sam-F     .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#Sam-En    .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
#NZM-4     .......... .......... .......... .......... .......... .......... .......... .......... .......... .......A..
#NZM-6     .......Y.. .......... .......... ..GK...... .A........ V......... .......... .......... S.....L... ..........
```

Table A.4 X gene amino acid sequences



Appendix B

Table B.1 Relevant background information on the collected hepadnaviridae genomes from Genbank.

All available human hepatitis B virus complete genomes were down-loaded from Genbank and analysed; Genotypes were assigned based on phylogenetic analysis with known genotypes quoted in the literature. Names and Origin are based on known place of viral infection. This was not always given in the primary literature and therefore was further researched, where this was not successful names where given in association with viral type eg, WHV for woodchuck sequences, or Genbank locus assignment. The Genbank locus and accession numbers are given as well as serotype (where known). The Isolate column gives the names assigned to the sequences by the authors in the primary literature. The references given are those where the origin of the sequence was first reported and does not always correspond to Genbank references.

G/T	Name	Locus	Accession #	Subtype	Isolate	Origin	Reference
F	Brazil	HBVADW4A	X69798	adw4	w4B	Rio de Janeiro	Naumann *et al.* (1993)
F	France 2	HHVBFFOU	X75658	adw4q-	Fou	France	Norder *et al.* (1994)
F	Colombia	HHVBF	X75663	adw4q-		Colombia	Norder *et al.* (1994)
E	Africa 2	HHVBE4	X75664	ayw4	Kou	Senegalese, Africa	Norder *et al.* (1994)
E	Africa	HHVBBAS	X75657	ayw4		West Africa	Norder *et al.* (1994)
D	Greece 1	HBVP2CSX	X97848	ayw	patient 2	Greece (wild type)	Alexpopulou *et al.* (1996)
D	Greece 3	HBVP3CSX	X97849	ayw	patient 3	Greece (mutant)	Alexpopulou *et al.* (1996)
D	German 2	HBVAYWGEN	Y07587	adw		Giessen, Germany	Stoll-Becker *et al.* (1997)
D	Greece 4	HBVP4PCXX	X80924	ayw	patient 4	Greece	Karayiannis *et al.* (1995)
D	Greece 5	HBVP5PCXX	X80926	ayw	patient 5	Greece	Karayiannis *et al.* (1995)
D	Greece 6	HBVP6PCXX	X80925	ayw	patient 6	Greece	Karayiannis *et al.* (1995)
D	Israel	HPBMUT	L27106		patient 2	Haifa, Israel	Hasegawa (1994)
D	Sardinia	HBVAYWMCG	X59795	ayw		Cagliari, Italy/Sardinia ?	Lai *et al.* (1991)
D	Poland 1	HBVGEN1	Z35716	ayw4	hb321	Poland	Pluciennicczak (1994). (UP)
D	Germany 5	HBVORFS	X72702	ayw3	C1005	Frieburg, Germany	Preisler-Adams *et al.* (1993)
D	Turkey	HPBHBVAA	M32138	?	a1-Turk	Turkey	Tong *et al.* (1990)
D	Latvia	XXHEPAV	X02496	ayw	pHB320	Latvia	Bichko *et al.* (1985)
D	HBVAYWC	HBVAYWC	X65257	ayw	patient C	Cagliari, Italy	Lai *et al.* (1992) (UP)
D	HBVAYWCI	HBVAYWCI	X65258	ayw	patient CI	Cagliari, Italy	Lai *et al.* (1992) (UP)
D	HBVAYWE	HBVAYWE	X65259	ayw	patient E	Cagliari, Italy	Lai *et al.* (1992) (UP)
D	HBVDNA	HBVDNA	X68292	ayw		Cagliari, Italy	Lai *et al.* (1992) (UP)

G/T	Name	Locus	Accession #	Subtype	Isolate	Origin	Reference
D	HBVPRES12	HBVPRES12	X85254			Cagliari, Italy	Lai *et al.* (1995) (UP)
D	HPBAYW	HPBAYW	J02203	ayw	*pHB320*	France/Europe	Galibert *et al.* (1979)
D	U.S.A. 2	U95551	U95551	ayw		U.S.A.	Rao *et al.* (1997) (UP)
C	Japan 1	HPBADRC	D00630	adr		Japan	Takemura *et al.* (1990)
C	Japan 12	HPE88A	D16666		E88	Nihon Uni. (Shinshu)	Uchida *et al.* (1995)
C	Japan 2	HPBADRM	D16665	adr	patient A4	Japan	Mukaide *et al.* (1992)
C	Japan	HPBH2B	D16667		H2	Shinshu Japan	Uchida *et al.* (1995)
C	Japan 20	D50517	D50517		HBV-ASA-EX1	Tokyo M&D Uni.	Asahina *et al.* (1996)
C	Japan 20-2	D50518	D50518		HBV-ASA-EX2	Tokyo M&D Uni.	Asahina *et al.* (1996)
C	Japan 21	D50519	D50519		HBV-ASA-EX3	Tokyo M&D Uni.	Asahina *et al.* (1996)
C	Japan 22	D50520	D50520		HBV-ASA-EX4	Tokyo M&D Uni.	Asahina *et al.* (1996)
C	Asia 9	HBVADR	V00867	adr	pHBR 330	Sth/E Asia / Japan	Ono *et al.* (1983)
C	Japan 13	HPBA11A	D50489		A11	Nihon University	Uchida *et al.* (1995)
C	Japan 16	HPBB4HST1	D23680		B4HST1	Jichi Medical School	Horikita *et al.* (1994)
C	Japan 17	HPBB5HKO1	D23682		B5-HBVKO1	Jichi Medical School	Horikita *et al.* (1994)
C	Japan 16-2	HPBC4HST2	D23681		C4-HBVST2	Jichi Medical School	Horikita *et al.* (1994)
C	Japan 17-2	HPBC5HKO2	D23683		C5-HBVKO2	Jichi Medical School	Horikita *et al.* (1994)
C	Japan 18	HPBC6T588	D23684		C6-TKB588	Jichi Medical School	Horikita *et al.* (1994)
C	Korea	HBVADRM	X14193	adr		Seoul Nat. Uni. Korean	Rho *et al.* (1989)
C	Japan 6	HPBETNC	L08805	adr	HT	Japan	Ogata *et al.* (1993)
C	E00188	E00188	E00188	adr			Kikuchi *et al.* (1984)
C	Japan 7	D28880	D28880		Nis 0859434	Fukuoka, Japan	Moriyama *et al.* (1994)
C	Japan 8	HEHBVAYR	X04615	ayr	pYRB259	Japan	Okamoto *et al.* (1986)
C	Japan 9	HPBADRA	M12906	adr	pHBV1-1	Japan	Kobayashi *et al.* (1984)
C	Japan24	HPBCG	D12980	adr		Japan	Mukaide *et al.* (1992)
C	Korea 2	HPBCGADR	M38636	adr		Korea	Kim *et al.* (1985)
C	S75184	S75184	S75184				Moriyama *et al.* (1994)
C	N/Caledonia	HHVBC	X75665	adrq-	HMA	New Caledonia	Norder *et al.* (1994)
C	F/Polynesia	HHVCCHA	X75656	adrq-	Cha	French Polynesia	Norder *et al.* (1994)

G/T	Name	Locus	Accession #	Subtype	Isolate	Origin	Reference
C	Japan 23	HBVADR4	X01587	adr4	pBRHBadr4	Japan	Fujiyama *et al.* (1983)
B	Asia 11	HBVCGINCX	X98073	adw	PF 2	Asia	Pult *et al.* (1997)
B	Asia 10	HBVCGINSC	X98072	adw 2	PF 1	Asia	Pult *et al.* (1997)
B	Asia 12	HBVCGWITY	X98077	adw	PF 6	Asia	Pult *et al.* (1997)
B	Japan 14	HPBA1HKK2	D23677		A1 HBVKK2	Jichi Medical School	Horikita *et al.* (1994)
B	Japan 19	HPBA2HYS2	D23678		A2 HBVYS2	Jichi Medical School	Horikita *et al.* (1994)
B	Japan 15	HPBA3HMS2	D23679		A3 HBVMS2	Jichi Medical School	Horikita *et al.* (1994)
B	Japan 3	D50521	D50521		HBV-ASA-FH1	Japanese	Asahina *et al.* (1996)
B	Japan 4	D50522	D50522		HBV-ASA-FH2	Japanese	Asahina *et al.* (1996)
B	China 4	HBVP4CSX	X97850	adw	patient 4	China (mutant)	Alexpopulou *et al.* (1996)
B	China 6	HBVP6CSX	X97851	adw	patient 6	China (wild type)	Alexpopulou *et al.* (1996)
B	Japan 10	HPBADW1	D00329	adw	pJDW233	Japan	Okamota *et al.* (1988)
B	Japan 11	HPBADW2	D00330	adw	pODW282	Okinawa, Japan	Okamota *et al.* (1988)
B	Indonesia	HPBADW3	D00331	adw	pAK66	Indonesia	Okamota *et al.* (1988)
B	Indonesia 2	HPBADWZ	M54923	adw		Indonesia	Sastrosoewignjo *et al.* (1987)
A	Germany 1	HBVXCPS	X70185	adw2.	A938	Freiburg/ Germany	Preisler-Adams (1993)
A	U.S.A. 1	HBVADW	X00715	adw	pHBv 933	U.S.A.	Ono *et al.* (1983)
A	E00120	E00120	E00120	adw			Onda *et al.* (1983)
A	E00192	E00192	E00192				Onda *et al.* (1983)
A	Germany 7	HBVA317	Z72478			Germany	Schories *et al.* (1996) (UP)
A	Germany 6	HBVA317MU	Z72479		A317	Froeburg, Germany	Schories *et al.* (1996) (UP)
A	Poland 2	HBVGEN2	Z35717	adw2		Poland	Plucienniczak (1994). (UP)
A	Asia 25	E01301	E01301	adw		Asia	Onda *et al.* (1987)
A	U.S.A. 3	S50225	S50225	adw2	U.S.A.	American	Wands *et al.* (1992)
	Asia 13	HBVDEFVP1	X98074	adw	PF 3	Asia	Pult *et al.* (1997)
	Asia 14	HBVDEFVP2	X98075	adw	PF4	Asia	Pult *et al.* (1997)
	Asia 15	HBVDEFVP3	X98076	adw	PF5	Asia	Pult *et al.* (1997)
	XXHEPA	XXHEPA	V01460	ayw	*Eco HBV DNA*	France	Galibert *et al.* (1979)
	E00010	E00010	E00010				Uiriamu *et al.* (1981)

G/T	Name	Locus	Accession #	Subtype	Isolate	Origin	Reference
	HBVPREX	HBVPREX	X52939	adr	hbvprex	W/Germany/China ?	Loncarevic *et al.* (1990)
	HPBADR1CG	HPBADR1CG	M38454	adr-1	pADR-1	Asia	Renbao *et al.* (1987)
	HPBADWZCG	HPBADWZCG	M57663	adw	hbv21/pFDW294	Philippines	Estacio *et al.* (1988)
	HBVADW2	HBVADW2	X02763	adw2		American	Velenzuela *et al.* (1980)
	chimpanzee	D00220	D00220	LSH	chimp.	London	Vaudin *et al.* (1988)
	gibbon	HBU46935	U46935	ayw 3g	gibbon	Thailand	Norder *et al.* (1996)
	WHV	OHVGD	M18752	WHV 7	wood chuck		Cohen *et al.* (1988)
	WHV.	OHCGC	M19183		wood chuck		Cohen *et al.* (1988)
	WHV	OHVHEPBA	J04514	WHV 8	wood chuck		Giranes *et al.* (1989)
	WHB	OHVCGA	J0244		wood chuck		Galibert *et al.* (1982)

Key

G/T = HBV Genotype

Name = name I have given associated with geographical origin correspond to Table 3.5

Locus = corresponds to Genbank assigned locus

Accession number = corresponds to Genbank assigned accession number

Subtype = defined by serotype

Origin = where viral infection occurred

UP = No publication was associated with the sequence at time of submission to Genbank

Appendix C
Neighbour-joining and maximum likelihood trees

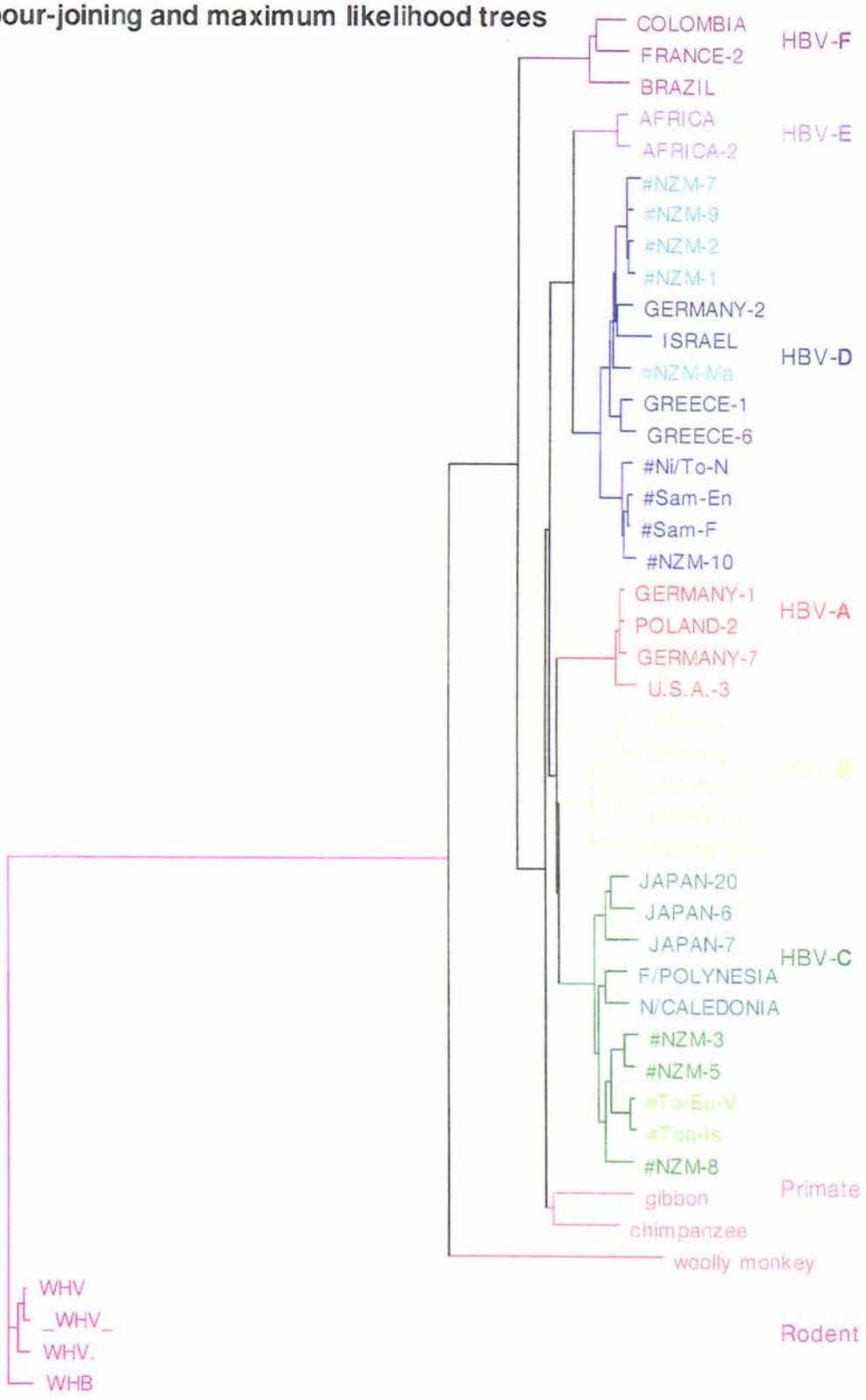


Figure C.1 Neighbour-joining tree of Pacific samples and 30 database hepadnaviridae sequences. The data corresponds to that used in the SpltsTree graphs (Figures 3.3 and 3.4). This tree, and the following, demonstrate more clearly than the SplitsTree graphs the order of divergence of the genotypes. Here HBV-**F** is placed as the first to diverge.

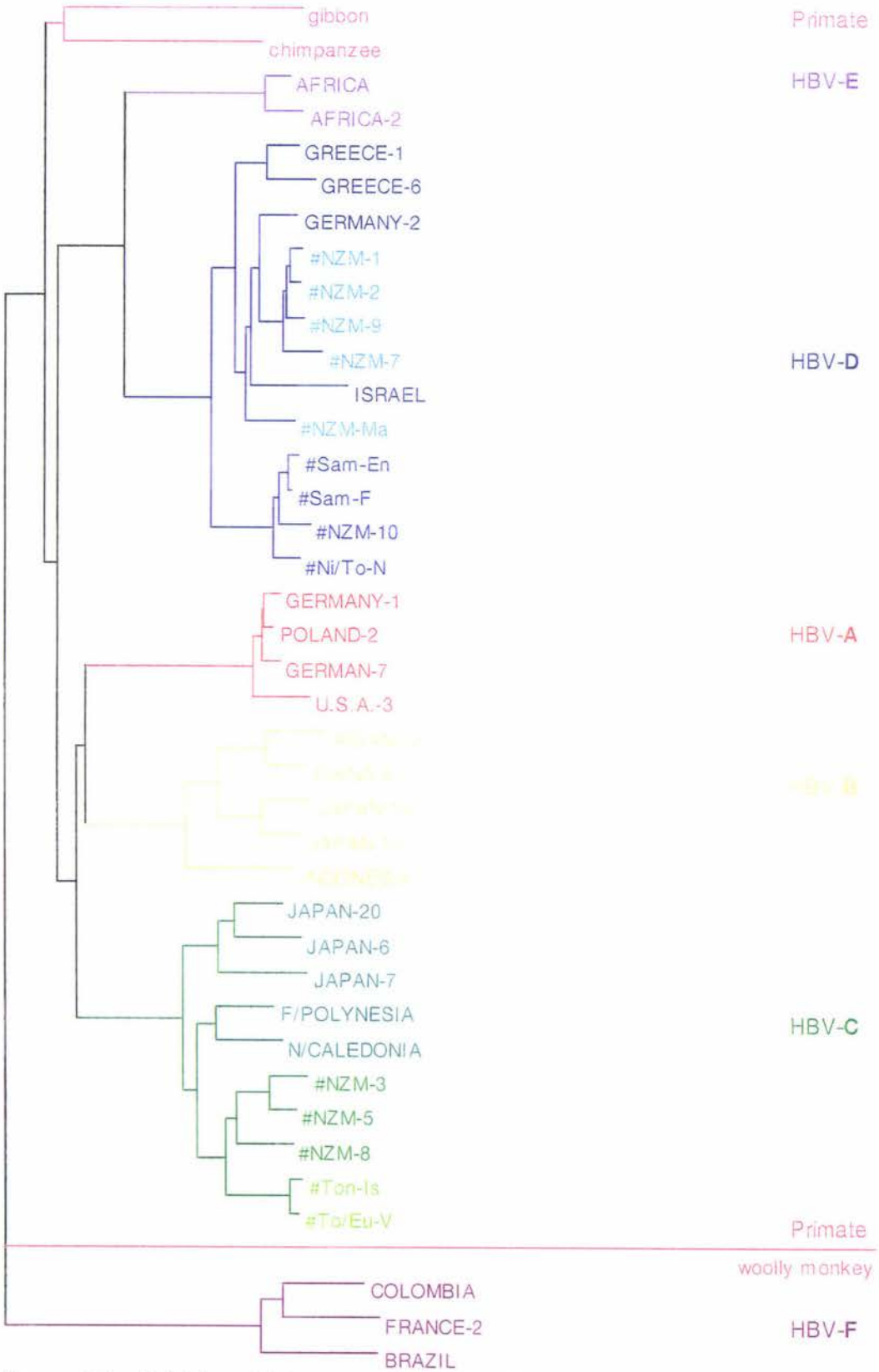


Figure C.2 Neighbour-joining tree of the Pacific samples and 22 database HBV sequences (rodent sequences are excluded from analysis). This tree clearly shows three clusters; cluster HBV-**A**, HBV-**B,** and HBV-**C**; cluster, HBV-**D** and HBV-**E**; and HBV-**F** forming its own early diverging cluster. This tree is similar to trees published by Norder *et al.* (1996) (see Figure 1.4).

Figure C.3 Maximum likelihood tree of the Pacific samples and 22 database sequences. Unlike the previous neighbour-joining trees here, HBV-**B** is placed as the first to diverge, the genotype clustering remains the same but the divergence pattern is different. This tree is similar to the maximum likelihood trees produced by Bollyky and Holmes (1998) that used a gamma distribution model, though no gamma distribution was used here (see Figure 1.6).

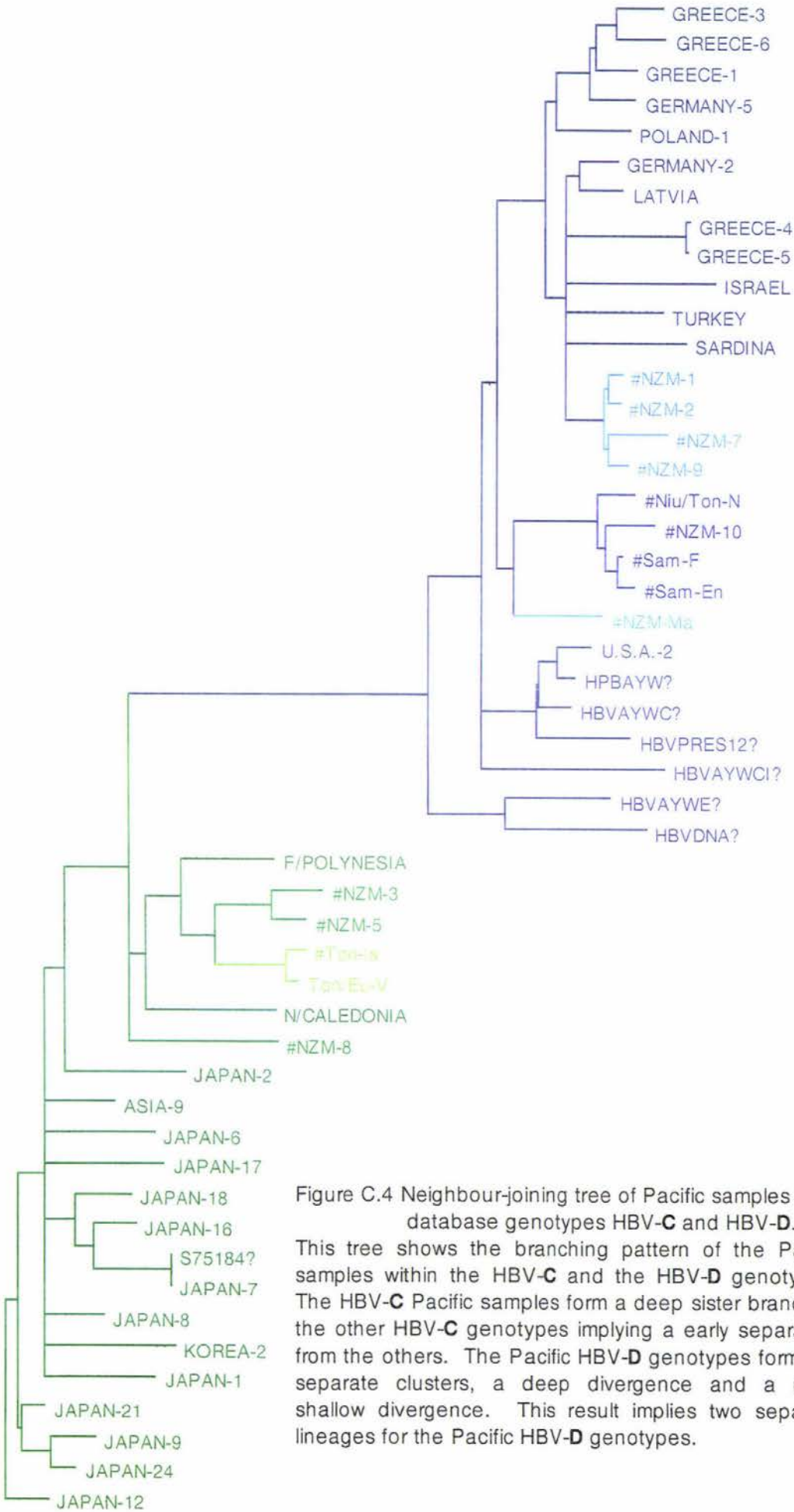


Figure C.4 Neighbour-joining tree of Pacific samples with database genotypes HBV-**C** and HBV-**D**.

This tree shows the branching pattern of the Pacific samples within the HBV-**C** and the HBV-**D** genotypes. The HBV-**C** Pacific samples form a deep sister branch to the other HBV-**C** genotypes implying a early separation from the others. The Pacific HBV-**D** genotypes form two separate clusters, a deep divergence and a more shallow divergence. This result implies two separate lineages for the Pacific HBV-**D** genotypes.

Appendix D

HLA-DPA1 paper submitted to Tissue Antigens

HLA-DPA1 sequencing based typing

Zlatibor M. Velickovic[1], Gabrielle L. Harrison[2], John M. Carter[1] and Rosalind P. Murray-McIntosh[2]

1 Department of Pathology, Wellington School of Medicine, Wellington South, New Zealand.

2 Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

Keywords: DPA1, HLA, Samoa, sequencing-based typing, Trobriand

Correspondence:

Zlatibor Velickovic

Department of Pathology, Wellington School of Medicine

[redacted]

Abstract:

The HLA-DPA1 gene coding for the α chain of class II DP antigen, is polymorphic with 13 alleles currently reported. The polymorphism of the HLA-DPA1 gene has been shown so far to influence the binding strength of antigenic peptides, play an important role in development of multiple sclerosis, and is associated with post-streptococcal acute glomerulonephritis. In this study we describe a reproducible and reliable method of sequencing-based typing for the HLA-DPA1 gene. We have designed M13 tailed PCR primers specific to the intron 1 and intron 2 sequences of the DPA1 gene for subsequent M13 dye primer sequencing. We have tested our method on 112 unrelated individuals of three ethnic groups Caucasian, Samoan and Trobriand, and 4 cell lines containing different DPA1 alleles obtained from the International Histocompatibility Workshop. The DPA1*0103 allele was found to be the most frequent variant in Caucasians (76%) similar to data already published. In comparison Trobrianders showed high allelic homogeneity (92%) while both Samoans and Trobrianders had a high frequency of the 02022 allele supporting the hypothesis of their Southeast Asian origin. HLA-DPA1 sequencing-based typing provides high resolution allele identification, good quality electropherograms, reproducibility and is suitable for routine clinical laboratory typing.

HLA genes constitute a highly polymorphic multigene system which codes for cell surface proteins presenting antigens to T lymphocytes. In the processes of targeting proteins originating in disease, HLAs play an essential role in the human immune system. In individuals, families and ethnic groups, particular polymorphisms in these genes have been clearly associated with increased or reduced susceptibility to specific diseases. The HLA-DPA1 gene codes for the α chain of class II DP antigen, the β chain of the complete antigen being provided by the HLA-DPB1 gene. Both genes are polymorphic with 13 DPA1 and 84 DPB1 alleles characterised so far (1, 2) and showing strong linkage disequilibrium in Caucasians. DPA1, the subject of our current study, has been investigated less than other HLAs so that consequences of its polymorphism are not yet well defined. Associations so far demonstrated for the polymorphism of DPA1 include influence on the binding strength of antigens (in consort with DPB1), an important role in the development of multiple sclerosis, and an association with post-streptococcal acute glomerulonephritis (3-5). Further associations with disease will undoubtedly arise from more thorough investigations. Importance of matching donor/recipient for DPA1 in bone marrow or solid organ transplantation is as yet unclear.

Sequencing-based HLA typing is the method that provides the highest resolution although it requires expensive equipment. Unlike PCR-SSP (5, 6) and PCR-SSO (8) methods, the sequencing based HLA typing has the capability of detecting new polymorphisms. A previously published HLA-DPA1 sequencing based typing method (9) used specific dye labeled primers for sequencing. However we use M13 tailed PCR primers which have the advantage of using widely available M13 dye primer sequencing kits and makeing the method more applicable for routine analysis.

In this study we describe a method for DNA sequence-based typing for the HLA-DPA1 gene. Like other Class II genes individual HLA-DPA1 alleles can be distinguished by analysis of the sequence of exon 2. We have designed M13 tailed PCR primers specific to the intron 1 and intron 2 sequences of the DPA1 gene for subsequent M13 dye primer sequencing. In order to test our method we have chosen 112 unrelated individuals from three different ethnic groups Caucasian, Samoan and Trobriand Islanders. In addition, 4 cell lines from the International Histocompatibility Workshop each containing a different DPA1 allele were also typed.

Caucasian and Pacific Island blood sample collection was approved by the Wellington Ethics Committee. Participant consent was obtained for the Trobriand samples by Dr Sciefenhoevel. DNA was extracted from whole blood using phenol-chloroform protocol and QIAmp blood kit (Qiagen) according to the supplied protocol. Four cell lines were purchased from the International Histocompatibility workshop through European Collection of Cell Cultures. They were: IHW 9010 AMAI. IHW 9031, BOLETH, IHW 9034SAVC, IHW 9076 T7526 specific to DPA1*0301, DPA1*0103, DPA1*0201and DPA1*0401 respectively. Unique primers specific to the intron 1 and 2 sequences of the HLA DPA1 gene were designed to amplify the intervening exon 2. In order to obtain high quality sequence and even peak distribution, dye primer cycle sequencing reactions were performed. M13 primer sequences were added to the 5' end of DPA1 specific primers. The forward M13 tailed primer was 5' TGT AAA ACG ACG GCC AGT GCT CCT TCT TCT TCC CCA TA 3' and the reverse M13 tailed primer was 5' CAG GAA ACA GCT ATG ACC CCC TCT CAT CCC TTC CAG TT 3' with the M13 sequence underlined.

PCR reactions in 50 µl volume contained the following: 50mM KCl, 10mM Tris-HCl pH9, 0.1% Triton X-100 (v/v), 2.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.2 µM of each primers, 1 Unit of Taq DNA Polymerase (Promega), 50 ng of DNA. PCR was performed using the following cycling conditions: 94^0C for 30 seconds and 68^0C for 1 minute for 30 cycles. PCR products were electrophorised on 1.5% agarose gel and visualised under UV light after ethidium bromide staining. The fragments were of expected length (483 bp). PCR reactions were washed with 400 _l of TE buffer purified of excess primers and concentrated to about 20 ng/µl using 30 KDa cut off spin columns (Pall Filtron). M13 dye primer cycle sequencing was performed using the PE-ABI Dye primer Cycle Sequencing Ready Reaction Kit with Ampli Taq FS DNA polymerase and samples were loaded on PE-ABI 377 automated DNA sequencer according to the manufacturer protocol. All samples which typed as homozygotes were subjected to subsequent PCR amplification with a set of primers specific to constant regions of exon 2 and 3 of the DPA1 gene in order to confirm the typing. The forward primer was 5' ATC CAG CGT TCC AAC CAC AC 3' and the reverse primer was 5' GGA AGA GGC TCT CAG CGA CA 3' with the PCR conditions and cycle parameters the same as above. PCR products were purified and sequenced using PE-ABI Big Dye terminator cycle sequencing according to the manufacturer protocol. Sequence files were analysed with Match Tools (PE-ABI) software by comparing them to a custom made allele library for the HLA DPA1 gene (available on request from the author). The software assigned an allele or alleles and possible allele combinations to each sample. Using MT Navigator (PE-ABI) software, the test sequences were aligned with the DPA1 exon 2 consensus sequence and polymorphic positions were re-examinedon the electropherogram. Any ambiguities were resolved and sequences were submitted to Match Tools analysis for final typing.

The DPA1 allele library for Match Tools was created using information on allele sequences from Anthony Nolan Bone Marrow Trust, HLA Informatics Group WWW page http://www.anthonynolan.com/HIG/data.html.

DNA from 29 Caucasian, 32 Samoan and 51 Trobriand Island blood samples were amplified by PCR and visualised on agarose gel. All of the samples amplified and produced unambiguous bands on the agarose gel of the expected base pair length (483 bp). Cycle sequencing performed on PCR products produced sequences that corresponded to the length of PCR fragments. Comparison with the consensus DPA1 sequence showed that the sequence information was correct. Six alleles DPA1*0103, DPA1*02011, DPA1*02012, DPA1*02021, DPA1*02022 and DPA1*0105 appeared in this group of samples as either homozygotes or heterozygotes (table 1). DPA1 sequencing-based HLA typing of the cell panels were the same as the originaly reported typings (see legend for figure 1). Typing of homozygote samples was confirmed by PCR amplification and subsequent sequencing as described in the methods.

The DPA1*0103 allele was found to be the most frequent variant in Caucasians (76%) (table1). This is consistent with published data (6, 8). A high frequency of homozygotes also was observed (76%). Samoan and Trobriander alleles also showed high homogeneity (47 and 92%), a lower frequency of the *0103 allele than Caucasians (42 and 10%) and a higher frequency of the 02022 allele (53 and 90% versus 15%). Trobrianders inhabit the Trobriand Islands of Melanesia and genetically appear to be of Southeast Asian origin (10). The high frequency of the 02022 allele supports this hypothesis as well as the high frequency and homozygosity of the DPB1*0501 allele in the same population (10); both 02022 and 0501 are considered to be variants typical of Asians (6, 11).

There is evidence that Samoans, as Polynesians, also migrated from a Southeast Asian origin (12, 13) and this too is supported by the high frequency of the 02022 allele. Samoan sample collection was accompanied with ancestry information given by the participants which showed 19% of European admixture; this may account for the levels of the 0103 allele seen here. Similar dilution of the DPB locus (specifically DPB1*0501 at 70%) (6) was observed in Western Samoans, whereas Trobrianders have shown strong restriction at DPB (90% typed as DPB1*0501) using the same sample base (10).

Sequencing-based typing of HLA genes is the method offering the highest possible resolution. Use of dye primer cycle sequencing chemistry produces electropherograms of uniform peak height; thereby identification of heterozygous positions is particularly easy because two peaks at the same position each produce approximately 50% of the height of the normal single base peak (as shown in figure 1). This method provides sequence of the complete exon 2 of the DPA1 gene. It is capable of distinguishing between the 13 DPA1 alleles identified so far. Sequencing-based HLA typing of DPA1 has advantages over existing methods in use. PCR SSP and SSO methods are limited by recognising only known alleles. Positioning the primers in the intron 1 and 2 regions provides the complete sequence of exon 2 thus eliminating the possibility of missing some polymorphic positions in the beginning or at the end of the sequence. Furthermore labeling such primers with fluorescent dyes is expensive and difficult to standardise in terms of creating special purpose-built mobility files. Fluorescent dye labeled M13 sequencing primers are commercially available with defined mobility files which makes them suitable for routine use. They are readily available and of lower cost then custom made labeled sequencing primers. Our sequencing-based HLA typing method produces sequences of good quality and reproducibility.

Although the cost of materials for this kind of typing is a little higher than for SSP and SSO typing, the time involved (1 day) makes the method suitable for routine application in tissue typing laboratories. It has been shown that DPA1 polymorphism plays a role in the strength of binding antigenic peptides (3), and this method of typing will assist further investigation in evaluating contributions from variants of this gene to the outcome of solid organ and bone marrow transplantation.

Acknowledgments

We thank the Health Research Council of New Zealand for a grant for this project and Dr Schiefenhöevel for the Trobriand Island samples arranged through Dr Robert Hickson. We appreciate the willingness of all those who contributed blood samples to the study.

References

1. Marsh S.G.E. HLA Class II region sequences, 1998. Tissue Antigens 1998: 51: 4ii: 467 - 507 2.

2. Marsh S.G.E. HLA Class I and II Sequence Alignments: Update December 1998 Anthony Nolan Bone Marrow Trust, HLA Informatics Group, Internet address: http://www.anthonynolan.com/HIG/data.html

3. Gaston JS, Goodall JC, Young JL, Young SP. Effect of polymorphism of the HLA-DPA1 chain on presentation of antigenic peptides. Hum Immunol 1997: 54: 40-7.

4. Yu M, Kinkel RP. Weinstock-Guttman B, Cook DJ, Tuohy VK. HLA-DP: a class II restriction molecule involved in epitope spreading during the development of multiple sclerosis. Hum Immunol 1998: 59: 15-24.

5. Sasazuki T, Kimura A, Ito Y. HLA-DP antigens and post-streptococcal acute glomerulonephritis. Acta Paediatrica 1996: 85: 916-8.

6. Aldener-Cannava A, Olerup O. HLA-DPA1 typing by PCR amplification with sequence-specific primers (PCR-SSP) and distribution of DPA1 alleles in Caucasian, African and Oriental populations. Tissue Antigens 1996: 48: 153-60.

7. Gilchrist FC, Bunce M, Lympany PA, Welsh KI, du Bois RM. Comprehensive HLA-DP typing using polymerase chain reaction with sequence-specific primers and 95 sequence-specific primer mixes. Tissue Antigens 1998: 51: 51-61.

8. Moonsamy PV, Klitz W, Tilanus MG, Begovich AB. Genetic variability and linkage disequilibrium within the DP region in the CEPH families. Hum Immunol 1997: 58: 112-21.

9. Rozemuller EH, Bouwens AG, van Oort E, et al. Sequencing-basedtyping reveals new insight in HLA-DPA1 polymorphism. Tissue Antigens 1995: 45: 57-62.

10. HagelbergE, , Kayser M, Nagy M, et al. Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Phil. Trans. R. Soc. Lond. B 1999: 353: 1-12

11. Stephens HAE, Brown AE, Dasnayance Chandanayingyong et al. The presence of the HLA class II allele DPB1*0501 in ethnic Thais correlates with an enhanced vaccine-induced antibody response to a malaria sporozite antigen. European J Immunol 1995: 25: 3142-7.

12. Hill AVS, Serjeantson SW, eds The colonization of the Pacific: a genetic trail. Oxford :Clarendon, 1989.

13. Murray-McIntosh RP, Scrimshaw BJ, Hatfield PJ, Penny D. Testing Migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. Proc Nat Acad Sci USA 1998: 95: 9047-9052.

Table1. HLA-DPA1 allele distribution in three ethnic groups

HLA-DPA1 types	Caucasians	Samoans	Trobrianders
0103/0103	20	6	3
0103/02011	1		
0103/02012	1		
0103/02022	2	15	4
02022/02022	2	9	44
02011/02022	2		
02011/02021		1	
0105/02022	1	1	

Figure 1. MT Navigator layout with the test sequence aligned to the DPA1 exon 2 consensus sequence and the corresponding electropherogram. Heterozygote bases are identified and verified visually (indicated by arrows).