LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Unravelling the evolutionary complexity of RNA viruses

Thesis submitted for the degree of

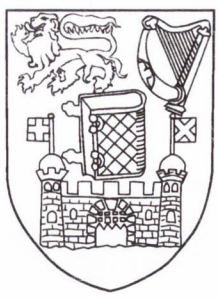
Doctor in Philosophy

September 2008

Damien Tully

Smurfit Institute of Genetics

Trinity College Dublin



TRINITY COLLEGE
10 AUG 2009
LIBRARY DUBLIN
THESIS
8768

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Damien Tully.

Signature ....

November 14, 2008

Summary

Viruses are a delicate group of organisms renowned for the serious diseases they inflict. They affect all cellular life and are subdivided into RNA-based and DNA-based viruses depending on the type of genomic nucleotide material they contain. Those that are RNA-based appear to contain more emergibility owing to their rapid rates of evolutionary change and are responsible for diseases such as Human Immunodeficiency Virus (HIV), Hepatitis C Virus (HCV) and Foot-and-Mouth Disease Virus (FMDV). The importance of dissecting the evolutionary complexity associated with such organisms is twofold: to improve our overall understanding of the mechanisms driving pathogenic viruses and to aid in predicting what genomic regions/structure motifs may be important for future antiviral therapeutics.

In the first part of this thesis, I focused on the deadliest animal virus of all - FMDV. Using an integrative evolutionary and computational approach I provide compelling evidence for heterogeneity of selection forces in shaping the evolution of the seven different FMDV serotypes. Such heterogeneity is further marked from structural and functional analysis that demonstrates a distinct pathway of evolution for those viruses confined to the African continent (SATs). Later, using a subset of sequences from the former single gene study I applied a newly devised coalescent method to estimate for the first time dates for the origin of the most recent common ancestor (MRCA) for FMDV and its respective serotypes. These results confirm a separate pathway of evolution for those SAT type viruses with a much older time of diversification for the remaining virus types (Euro-Asiatic). This older time of diversification confirms previous historical observations and suggests, although speculative, that Europe acted as a hub for the disease from where it transmitted elsewhere due to exploration and trading routes.

The dichotomy between virus type and genome regions was further examined when the complete genome of FMDV was explored using a large global dataset. In this work I examined how different genomic regions were under different selective constraints and postulated that a selection-drift balance accounts for the emergence of variability within FMDV isolates. This changing balance has led to the differential accumulation of slightly deleterious mutations and compensatory changes between genomes and serotypes. I also postulated that these mutational dynamics help account for the diverse epidemiological patterns witnessed by the different FMDV serotypes.

In the second part of this work I utilised two major viruses of great human significance - HIV and HCV. In this work I dissected the coevolutionary patterns operating within the functionally important envelope gene for the group of viruses responsible for the current pandemic. The use of this gene leads us to correlate many of the amino acids whose evolution is inextricably linked over the entire HIV-1 group M pandemic with significant biological functions. These biological

functions range from CD4 binding to glycosylation. In addition those not found to have a previously determined function may represent novel functional dependencies whose evolution has been functionally/structurally constrained over the evolution of group M. This study also examined the recent evidence demonstrating that a single amino acid residue in the functional *gag* gene was required for the adaptation of HIV-1 to humans. Using sequence data from the 3 groups of HIV-1 (M, N and O) and those infecting chimpanzee species (SIVcpz) I illustrated that the same functionally important amino acid acted in concert with a number of surrounding amino acids thereby facilitating a network of functional dependencies that most likely served in the adaptation of HIV-1 to its human host.

Finally I investigated the evolutionary nature of Hepatitis C in a large dataset of complete genomes representing all six genotypes and at this time all known subtypes. It is widely conceived that genetic drift is the ultimate source of genetic variation among viral isolates of different type. However, no single study to date has ever examined this relationship among all the genotypes and furthermore the extent to which recombination and positive selection act remain largely unknown. Therefore it is important to detect and report the presence of both evolutionary processes in order to bring forth new strategies for public health control and clinical treatment. Hence, in this study I quantified the factors that vary in a subtype-specific manner and explored the main evolutionary forces shaping the genome of HCV across six genotypes. I found interesting connections between the analysis of slightly deleterious mutations and synonymous sites all which definitely incriminate genetic drift in accounting for the observed sequence diversity.

Acknowledgments

My sincere thanks go to Mario for being a supervisor and mentor. For all the opportunities to travel and meet the weird and wonderful scientists but above all to learn. Thank you for providing me with a stimulating and enjoyable work environment. For being tolerant of me and enthusiastic with my work. To everyone in the lab past and present: Christina, Dave, Fran, Orla, Paco, Simon, Tom, Valentin and Xiaowei, thanks for putting up with me and for acting as sounding boards! A special word of appreciation to Jenny and Valentin. Firstly, Jenny for all your assistance, from your fine critiquing to simply maintaining my sanity. Secondly, Valentin for answering endless computer questions beyond my knowledge.

To my friends I owe a huge thank you for all your patience and being very understanding for my lack of presence in recent times, Andy, Garbhan, Gayle, James, Luke, Michelle, Noreen, Redser and Phil. Thanks for all the greats nights full of endless fun and laughter. In particular Michelle you have been an absolute star and a great friend always willing to help and cheer me up. A special mention to all the members and instructors of Loreto College Kenpo Karate club for providing an environment where I could literally switch off. A sanctuary like place where I could escape to and sweat like a pig but yet still do so with enjoyment surrounded by close friends.

To the McManus family, Finnonula, Jim and Gra I am grateful to you all for your support, encouragement and brutal honesty. In addition I say thanks for all the yummy food that gave me the fuel necessary to function properly and for making me a part of your family.

I owe a serious level of gratitude and I am indebted to my family. My parents, Cathal and Ger, my brothers, Jonathan and Chris and also the elder lemon my grandad Jack. Thank you for trying to understand what I do and for always offering support filled with words of wisdom. You always backed me in whatever I wanted to do and all I can say is thank you very much for everything

Finally, but certainly not least the woman who has put up with me over the years - Meadhbh. A woman of immense strength and belief you have gone beyond and above the calls for me and helped in anyway possible from cooking to proofreading. I offer my profound gratitude to you for your endless level of support, tolerance, encouragement, understanding and above all love. Although it is impossible to describe what you mean to me and how you have helped me in all walks of life. In the words of the immortal song "your simply the best" and have surpassed any expectations that one could ask from a loved one. Thank you for your belief.

It is a cursed evil to any man to become as absorbed in any subject as I am in mine.

Charles Darwin

Contents

1 Introduction **1**

1.1 Viruses, classes and classification 1

1.2 RNA viruses 3

1.2.1 Features of RNA viruses 4

1.3 Molecular epidemiology of RNA viruses 5

1.3.1 Inferring epidemiological history 5

1.3.1.1 Coalescent Theory and Methodologies 6

1.3.2 Application of Forensics to RNA viruses 9

1.3.3 Phylodynamics 10

1.4 Rates of evolutionary change 11

1.4.1 Mutation Rates 11

1.4.1.1 Measuring mutation rates 12

1.4.2 Substitution rates 12

1.4.2.1 Measuring substitution rates 14

1.4.3 Causes of high evolutionary rates 14

1.4.4 Consequence of evolutionary rates 15

1.5 Evolutionary mechanisms associated with viral genomes 16

1.5.1 Recombination 17

1.5.1.1 Methods to detect Recombination 18

1.5.2 Natural Selection 19

1.5.2.1 Methods to detect positive selection 21

1.6 Epistasis and coevolution 23

1.6.1 Epistasis 23

1.6.2 Coevolution 24

1.7 Debate surrounding RNA virus evolution 25

1.7.1 The Model of quasispecies 25

1.7.2 The relevance of quasispecies to RNA viruses 26

1.7.3 Survival of the flattest . . . 27
1.8 Summary, goals and achievements . . . 28
1.8.1 Human Immunodeficiency Virus (HIV) . . . 29
1.8.2 Hepatitis C Virus (HCV) . . . 30
1.8.3 Foot-and-Mouth Disease Virus (FMDV) . . . 30
1.8.4 Achievements . . . 31

2 Selection shifts in Foot-and-Mouth Disease Virus serotypes 33
2.1 Abstract . . . 33
2.2 Introduction . . . 34
2.3 Materials and Methods . . . 36
2.3.1 Sequence data . . . 36
2.3.2 Sequence alignments and phylogenetic tree inference . . . 36
2.3.3 A first test of changes in selective constraints using neural networks . . . 36
2.3.4 Characterising the main selective constraints in FMDV serotypes . . . 37
2.3.5 Testing Functional divergence between serotypes . . . 38
2.3.6 Identifying coevolution between structural/functional regions in VP1 . . . 39
2.4 Results . . . 40
2.4.1 Heterogeneous patterns of adaptive evolution between serotypes . . . 40
2.4.2 Evidence for changes in selection constraints among FMDV serotypes . . . 43
2.4.3 Functional divergence between FMDV serotypes . . . 44
2.4.4 Intra-molecular co-variation generates FMDV immune-escape mutants . 45
2.5 Discussion . . . 48
2.5.1 Heterogeneous selective constraints and the emergence of immune-escape mutants in FMDV serotypes . . . 48
2.5.2 Host-cell receptor recognition and antigenicity may be linked by positive selection and coevolution . . . 51
2.5.3 A geographical-specific pattern of adaptive coevolution in FMDV . . . 52
2.5.4 A geographical-based pattern of functional divergence between FMDV serotypes . . . 53
2.5.5 Virus variability and survival . . . 53
2.5.6 Sampling bias . . . 53
2.6 Acknowledgements . . . 53

3 The tale of a modern Animal Plague: Tracing the evolutionary history and timescale for Foot-and-Mouth Disease Virus **54**

3.1 Abstract . . . 54

3.2 Introduction . . . 55

3.3 Material & Methods . . . 57

3.3.1 Viral Sequence Data . . . 57

3.3.2 Sequence Analysis . . . 57

3.3.3 Bayesian MCMC evolutionary analyses . . . 58

3.4 Results . . . 58

3.4.1 Estimation of nucleotide substitution rates . . . 61

3.4.2 Divergence times of FMDV serotypes . . . 63

3.4.3 Demographic history of FMDV . . . 63

3.5 Discussion . . . 64

3.6 Acknowledgments . . . 68

4 Shifts in the Selection-Drift Balance Drives the Evolution and Epidemiology of Foot-and-Mouth Disease Virus **69**

4.1 Abstract . . . 69

4.2 Introduction . . . 70

4.3 Material and Methods . . . 72

4.3.1 Genome Sequences and Phylogenetic Analyses . . . 72

4.3.2 Identifying Adaptive Evolution in FMDV Genomes . . . 73

4.3.3 Identifying Constraints at Synonymous Sites . . . 74

4.3.4 Identifying and Quantifying Slightly Deleterious Mutations . . . 74

4.3.5 Identifying Compensatory Mutations in the FMDV Proteome . . . 75

4.4 Results . . . 76

4.4.1 Different Evolutionary Forces among Genomic Regions of FMDV Serotypes . . . 76

4.4.2 Relaxed Selective Pressures at Structural Capsid Proteins in SAT Serotypes . . . 78

4.4.3 Differential Fixation of Slightly Deleterious Mutations among FMDV Serotypes . . . 78

4.4.4 Compensatory Mutations Generate Immune-escaping Mutants in SAT Serotypes . . . 82

4.5 Discussion . . . 84

4.5.1 Diversifying Selection Has Driven the Evolution of FMDV Genomes . . . 85

4.5.2 Varying Selective Constraints at Synonymous Sites between FMDV Genomic Regions . . . 86
4.5.3 High rates of fixation of SDMs at SAT serotypes . . . 86
4.5.4 Compensatory Mutations Generate Immune-escaping Mutants in SAT Serotypes . . . 88

5 A study of the coevolutionary patterns operating within HIV-1 using the *gag* and *env* gene 90
5.1 Abstract . . . 90
5.2 Introduction . . . 91
5.3 Materials and Methods . . . 94
5.3.1 Taxon Selection . . . 94
5.3.2 Analysis of intra- and inter-protein molecular coevolution . . . 95
5.3.3 Mapping significant amino acid residues onto 3D protein structures . . . 97
5.4 Results . . . 98
5.4.1 Coevolution of a functionally important residue in *gag* . . . 98
5.4.2 Coevolution analysis in *env* gene belonging to group M . . . 100
5.4.2.1 Coevolution within the gp120 V3 loop and proposed co-receptor binding domains . . . 100
5.4.2.2 Networks of coevolving amino acids between gp120 and CD4 . . . 103
5.4.2.3 Coevolution within gp120 moving domains . . . 109
5.4.2.4 Inter gp120-gp41 coevolution . . . 109
5.5 Discussion . . . 112
5.5.1 Co-adaptation of *gag* to its human host . . . 112
5.5.2 Coevolutionary patterns operating within *env* . . . 113
5.5.3 The distribution of coevolving residues throughout the *env* gene . . . 115
5.5.4 Coevolution within the gp120 V3 loop . . . 116
5.5.5 Coevolution networks demonstrate the complexity of evolution operating within *env* . . . 117
5.6 Acknowledgements . . . 119

6 Understanding the genetic diversity of HCV 120
6.1 Abstract . . . 120
6.2 Introduction . . . 121
6.3 Materials and Methods . . . 124
6.3.1 Genomic Data . . . 124

6.3.2 Detection of selective pressures at single amino acid site and individual lineages . . . 124
6.3.3 Recombination analysis . . . 125
6.3.4 Identifying Constraints at Synonymous Sites . . . 126
6.3.5 Identifying and Quantifying Slightly Deleterious Mutations . . . 126
6.3.6 Codon Usage Analysis . . . 127
6.4 Results . . . 127
6.4.1 Heterogeneous selective constraints . . . 127
6.4.2 Recombination . . . 130
6.4.3 Homogeneity of synonymous sites . . . 135
6.4.4 Codon adaptation index of genes and genotypes . . . 139
6.4.5 Heterogeneity of SDMs . . . 139
6.5 Discussion . . . 140
6.5.1 Natural selection as the architect of HCV genotype evolution? . . . 141
6.5.1.1 Previous Studies highlighting Adaptive Evolution in HCV . . . 142
6.5.2 Recombination as a rare or absent process in HCV isolates . . . 143
6.5.3 Genotype diversity . . . 145
6.6 Acknowledgements . . . 146

7 General discussion **147**

Bibliography **152**

List of Figures

1.1 Schematic representation of the general replication cycle for viruses 2
1.2 The basic principle of coalescent theory. 7
1.3 Representation of tree shapes under different phylodynamic processes. 11
1.4 The average range of substitution rates for viruses. 14
1.5 A simple test to detect recombination 20
1.6 Types of mutations in protein coding DNA 22
1.7 Schematic drawing of the survival of the flattest effect 28

2.1 Number of functional and structural important sites in each serotype. 44
2.2 Analysis of functional divergence between FMDV serotypes. 45
2.3 Analysis of the distribution of co-evolving sites in the different serotypes. 46
2.4 Coevolution analyses in the VP1 protein from FMDV serotype O. 47
2.5 Coevolution analysis in the VP1 protein from FMDV serotype SAT1 49
2.6 Cumulative mean Poisson-corrected pairwise distance along the sequence alignments of FMDV 51

3.1 Maximum Clade Credibility Tree based on 236 FMDV gene sequences 60
3.2 Bayesian Skyline Plot estimated from the entire dataset of FMDV isolates sampled between 1932 and 2001 64
3.3 Hypothesis on the spread of FMDV from ancient times. 67

4.1 Distribution of the percentage of codon regions under adaptive evolution along the genome of FMDV for serotypes 77
4.2 Genomic distribution of nucleotide substitutions per synonymous sites (d_S) . . . 79
4.3 Phylogenetic distribution of the percentages of slightly deleterious mutations (SDM) in each serotype 81
4.4 Comparison of the percentage of slightly deleterious mutations 82

5.1 Amino acid residues coevolving in the *gag* encoded matrix protein 99

5.2 Coevolving pairs of amino acid residues over the complete *env* gene 102
5.3 Pairwise coevolving residues observed within the gp120 V3 loop 104
5.4 The CD4 coevolution network 106
5.5 The glycosylation coevolution network 108
5.6 Inter-protein coevolution between gp41 and gp120 111

6.1 Constructed phylogeny of the core gene for HCV isolates 131
6.2 Constructed phylogeny of the E1 gene for HCV isolates 132
6.3 Constructed phylogeny of the E2 gene for HCV isolates 133
6.4 Constructed phylogeny of the NS4B gene for HCV isolates. 135
6.5 Maximum likelihood phylogenetic tree depicting the topology of HCV genotypes for the Core gene. 136
6.6 Maximum likelihood phylogenetic tree depicting the topology of HCV genotypes belonging to the E2 gene 137
6.7 Distribution of synonymous nucleotide substitutions (d_S) in the polyprotein opening reading frame of HCV for all six genotypes. The medians for each genotype are show in yellow in all plots. 138
6.8 Codon adaptation index for each gene and its respective genotype 139
6.9 Comparison of the percentage of slightly deleterious mutations between HCV genes and genotypes 140

List of Tables

2.1 Amino acid sequence regions undergoing positive selection 42
2.2 Hemi-matrix of the number of sites under functinal divergence type 1 among FMDV serotypes 45

3.1 Bayesian estimates of divergence times and evolutionary parameters for FMDV serotypes 62

4.1 Fractions of slightly deleterious mutations in FMDV genomes and serotypes . . . 80
4.2 Distribution of the percentage of SDMs in the different phylogenetic levels in SAT types compared to Euro-Asiatic serotypes 80
4.3 Complete list of all compensatory mutations detected in each serotype 83

5.1 Mean nucleotide pairwise substitutions per site for the different subtypes used in this study 96
5.2 Residues detected as coevolving with site 30 and correlated by hydrophobicity (Hydro) and molecular weight (MW). 98
5.3 Details of the coevolving pairs whose coevolution was correlated by hydrophobicity, molecular weight, or both 101
5.4 Residues in the V3 Loop and proposed CCR5-Binding domain that coevolve with residues involved in CD4 binding 105
5.5 Coevolving amino acid residues that exhibit marked differences in their mean pairwise distances between the liganded HIV gp120 and the unliganded SIV gp120 structures 110
5.6 Residues within the proposed gp120 binding hydrophobic patch of the gp41 ectodomain that coevolve 111

6.1 Likelihood ratio test for the comparsion between codon based models 128
6.2 Recombination events inferred using a number of different methods contained within the RDP software package 134

Chapter 1

Introduction

Preface

In the first part of this introduction I describe some background on the concepts of molecular evolution and their application to RNA virus study systems. In the second part I address the state of the field in evolutionary genetics in determining the rates of evolutionary change in viruses and the challenges and opportunities. This section also outlines the research chapters that investigate the impact of such evolutionary mechanisms in important human and animal RNA viruses.

1.1 Viruses, classes and classification

Viruses are small, infectious, obligate intracellular parasites found in all cellular forms of life, from bacteria to chordates. They are the last major group of organisms to be described and may represent the last and, indeed, the broadest organismal frontier. One of the notable features of viruses is how greatly they differ in terms of their genome organisation and of the modes of their genome expression. Viruses whose genome is RNA rather than DNA, as in all cellular organisms, utilise continuous genetic change as they explore sequence space to improve their fitness and adapt to the changing environment of their hosts. Such variation is intimately linked to their disease-causing potential. The association between viruses and their hosts range from ephemeral once-off visits to chronic fatal associations.

At its simplest, a virus is formed by a small number of nucleic acid molecules protected by a closed protein shell, which enters a host cell as a prerequisite to replication. The entry of a virus into a host cell is mediated by the interaction of virus proteins with corresponding host cell receptors. Although entry is only the first challenge that a virus faces during the replication process, it also must propagate its own genome in the presence of an immune response from the host organism. Once an invading body is recognised the host will send out specific and non-

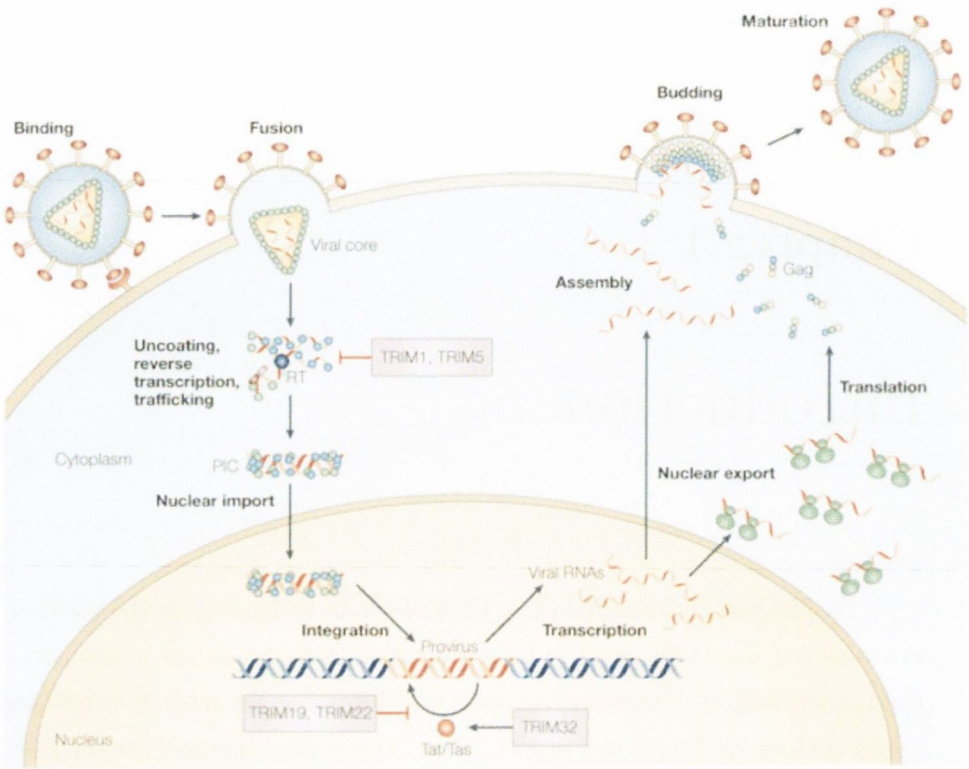


Figure 1.1: Schematic representation of the general replicative cycle for viruses. In this case the cycle is shown for retroviruses such as HIV. Reprinted with permission from Macmillan Publishers Ltd: Nature Genetics (Nisole et al., 2005).

specific responses to protect the body from serious threats. At this point the virus must overcome these defence mechanisms and gather all the necessary replication machinery to produce copies of its genome and virion components. The ultimate goal is the packaging of newly produced genomes within virions, exodus from the cell and invasion of new host cells. Eventually the descendants will leave the body in search of the next susceptible host (Figure 1.1).

Accounting for the basic distinction between nucleic acid types is a broad virus classification system. Viruses whose genomes are composed of DNA (DNA viruses) and those where the genomic nucleic acid comprises RNA (RNA viruses) are divided with the latter including retroviruses that possess genomic material comprised from RNA but replicate via a DNA stage facilitated by reverse transcription. Although the earliest classification encompassing all viruses was based mainly on about a dozen phenotypic characteristics including morphology, nucleic acid type, mode of replication and the type of the disease caused. Today the most widely accepted system is that of the Baltimore classification scheme which places viruses into one of seven groups depending on a combination of their nucleic acid (DNA or RNA), strandedness (single-stranded (ss) or double-stranded (ds)), and method of replication (Baltimore, 1971). Classifying viruses in this manner leads to viruses been placed in one of the seven following groups (Baltimore,

1971):

- I: dsDNA viruses (e.g. Adenoviruses, Herpesviruses, Poxviruses)
- II: ssDNA viruses (+)sense DNA (e.g. Parvoviruses)
- III: dsRNA viruses (e.g. Reoviruses)
- IV: (+)ssRNA viruses (+)sense RNA (e.g. Picornaviruses, Togaviruses)
- V: (-)ssRNA viruses (-)sense RNA (e.g. Orthomyxoviruses, Rhabdoviruses)
- VI: ssRNA-RT viruses (+)sense RNA with DNA intermediate in life-cycle (e.g. Retroviruses)
- VII: dsDNA-RT viruses (e.g. Hepadnaviruses)

Many of the emerging diseases that threaten humans are caused by RNA viruses, a trait not by chance (Cleaveland et al., 2001; Woolhouse, 2002; Cleaveland et al., 2007). The clear bias in which RNA viruses cross the species barrier more often than DNA based viruses is due to the process of co-speciation between DNA viruses and viable hosts that extend millions of year. Undoubtedly it is the inherent high mutation rates in RNA viruses that provides them with the capacity to adapt very quickly to changing host environments and to overcome host barriers to spread. Although other factors may be responsible, and will be discussed at a later stage. Owing to this, attention throughout this work will be given to those diseases caused by RNA viruses. However references will be made with respect to DNA viruses to highlight the dichotomy that exists between the two groups. Furthermore it will be demonstrated that RNA viruses inherently harbour more evolutionary potential and as such more "emergibility" than DNA viruses.

1.2 RNA viruses

At present, there are complete genome sequences for $\sim$ 600 species of ss positive-sense viruses (`http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239`), which include poliovirus, SARS coronavirus, FMDV, dengue and the 'common cold' rhinovirus. The remaining RNA virus groups are the ss negative sense viruses of which there are $\sim$ 120 completely sequenced genomes which include influenza, measles, rabies and Ebola virus; the ds viruses of which there are $\sim$ 110 genomes which include the veterinary pest bluetongue and rotavirus which causes diarrhoea. Finally the retrotranscribing viruses contain $\sim$ 100 complete genomes namely HBV and the retrovirus implicated in AIDS. So as one can see the wealth of whole genome sequences available on public databases such as GenBank (Ghedin et al., 2005) provides us with copious grist to perform evolutionary genomic studies. This will allow us to uncover the evolutionary

mechanisms that allow viruses to infect new hosts and establish self-sustaining transmission chains. This coupled with the knowledge that RNA viruses feature predominantly in the list of the most serious infectious diseases makes it an attractive organism to study. In fact, the second and sixth biggest killers worldwide are RNA viruses (HIV and measles, respectively), and several RNA viruses have been implicated to the first and third biggest killers: lower respiratory infections and diarrhoea, respectively (WHO, 2004). By contrast, none of the known DNA viral species appear on the list within the top 30. .

1.2.1 Features of RNA viruses

The interest in studying the evolution of RNA viruses is not only motivated by the need to develop new rational antiviral therapies. They also serve as excellent experimental models for the study of evolution in general. RNA viruses change so rapidly that it is possible to watch spatial and temporal patterns unfold in "real time", a scale that is usually not visible in other organisms. The rapidity of RNA virus evolution is as a result of a number of general properties.

(a) RNA virus replication is a very error prone process due to the lack of proof-reading mechanism in their RNA-dependent RNA polymerase (Steinhauer et al., 1992). As a consequence they exhibit the highest mutation rate for any group of organisms in the order of 10^{-4} to 10^{-5} misincorporations per nucleotide per round of copying (Drake and Holland, 1999; Mansky, 2000; Crotty et al., 2001). This means that approximately one nucleotide error per genome is made during every round of replication (Malpica et al., 2002; Drake, 1993).

(b) Genomes are small, ranging from only 3 kb (e.g. levivirus) to the largest (coronavirus) at 30 kb, with a typical size of 9 kb. It is generally thought that the small size is intimately associated with the last property because high mutations rates are theoretically supposed to limit genome size (Holmes, 2003). This is due to the so called error threshold as most mutations are deleterious. Hence even the fittest viral genomes are unable to reproduce without incurring lethal mutations with population size decreasing to extinction (Eigen, 1971). Therefore the readily adaptable nature of RNA is somewhat paradoxically constrained by the nature of their small size. Such constraint is consistent with the observations that some RNA viral genomes show signs of compression with the widespread use of overlapping reading frames to encode multiple functions for some nucleotides (Belshaw et al., 2007). In fact, RNA viruses were given the metaphor "as restless beasts pacing a small cage" to highlight their rapid adaptive exploitation of a limited array of possibilities (Belshaw et al., 2008).

(c) Their immense population sizes several orders of magnitude larger than any population

size for DNA based organisms such that the number of viral particles in a given infected organism may be as high as 10^{12} a direct result of explosive replication (Domingo and Holland, 1997).

(d) Such population sizes are directly related to short viral generation times where a single infectious particle can produce, on average, 100,000 copies in 10 hours (Domingo and Holland, 1997).

Together, these factors produce rates of nucleotide substitution that are higher than those in eukaryotes and DNA viruses (Jenkins et al., 2002). Finally, although viruses lack a fossil record, their evolution can often be recorded over the timescale of human observation. One of the most important advances in the study of viral dynamics has been the birth of molecular epidemiology and the development of coalescent-based methodologies to estimate rates of nucleotide substitution and to infer population dynamics.

1.3 Molecular epidemiology of RNA viruses

The term molecular epidemiology was first introduced to the study of infectious disease in the early 1970s (Kilbourne, 1973). The term is widely used in many different ways by different people but the term itself partly comes from the Greek with epidemiology derived from epidemic meaning "upon the people", whereas molecular refers to molecular biology - the analysis of molecules, usually nucleic acid and proteins. Therefore, a logical interpretation of the term is a marriage of the two, to use gene sequence data in the study of health and disease determinants in human populations (Hall, 1996). Although other definitions have also been discussed (Foxman and Riley, 2001). In reality, this young and still emerging discipline is providing new fascinating insights into the biology, origin and spread of microbial pathogens (Holmes, 1998b,a; Carrington et al., 2005; Lemey et al., 2003; de Oliveira et al., 2006; Robbins et al., 2003; Nelson et al., 2008; Rambaut et al., 2008; Deurenberg and Stobberingh, 2008; Pybus et al., 2001). In this section, I will briefly discuss how we can reconstruct epidemiological history from sequence data of infected individuals and show, using examples, how this type of analysis is far from a standard part of exploratory sequence analysis but, in some cases, it can literally be a matter of life and death and several studies have shown this fact.

1.3.1 Inferring epidemiological history

While the use of phylogenetic trees have primarily been used for taxonomic purposes to determine how closely organisms are related to one another it now plays an increasing role in molecular epidemiology where it provides us with information about the extent of genetic diver-

sity within populations. While phylogenetic estimation and the stages of a molecular analysis is outside the remit of this discussion, excellent works reviewing both are available (Perez-Losada et al., 2007; Holder and Lewis, 2003; McCormack and Clewley, 2002; Whelan et al., 2001). The traditional phylogenetic methods of estimating trees have a number of limitations associated with them. Firstly, the evolutionary tree is in reality an assumption of how evolution takes place, not a truism. Secondly, such an approach makes little sense for questions that involve more complicated demographic scenarios. In such cases, if recombination between pathogens is widespread then the relationships between sequences should not be estimated, as a phylogenetic tree generally depicts strictly bifurcating lineages. Therefore, a tree representation fails to accurately portray a reasonable genealogy. Finally, the evolutionary history of viral species or genes is simply equated from sampling of a single sequence of each species under study.

By contrast, such limitations have produced gene genealogical trees that are graphically similar but fundamentally very different (Rosenberg and Nordborg, 2002). Genealogies do not estimate trees but depict the history of genetic polymorphisms segregating in contemporaneous populations. Furthermore, the genealogical approach has none of the limitations of the phylogenetic method and provides a coherent statistical framework to consider recombination and selection. Moreover, using a stochastic process known as the coalescent we can infer demographic processes from analysis of genetic polymorphisms.

1.3.1.1 Coalescent Theory and Methodologies

Coalescent theory has revolutionized molecular population genetics over the past 20 years. Using this theory we simulate the genealogy of the sample going back in time until the most recent common ancestor (MRCA) is found (Figure 1.2). It was modeled independently by several authors in the early 1980s (Hudson, 1983b,a; Tajima, 1983) although its definitive discovery is testament to Kingman (Kingman, 1982, 2000). The underlying idea is that whenever two lineages pick the same parent, their lineages coalesce and eventually all lineages coalesce into a single lineage termed the MRCA of the sample.

As a tool, the coalescent serves very useful as a mathematical modelling tool that can be used to derive estimates of population parameters, as a simulation tool for various hypothesis testing and lastly for exploratory data analysis. The development of powerful analytical tools under the coalescent framework has led to important advances in the study of phylogeography (Martinez-Solano et al., 2007; Weinstock et al., 2005; Hayward and Stone, 2006; Grazziotin et al., 2006), molecular ecology (Liu et al., 2008; Storz et al., 2002; Flanagan et al., 2004) and anthropology (Reich and Goldstein, 1998; Shapiro et al., 2004; Edwards et al., 2007; Rogaev et al., 2006). Another area where coalescent approaches has had a marked impact is the evolutionary genetics and epidemiology of infectious disease (Moya et al., 2004; Duffy et al., 2008). Indeed, RNA

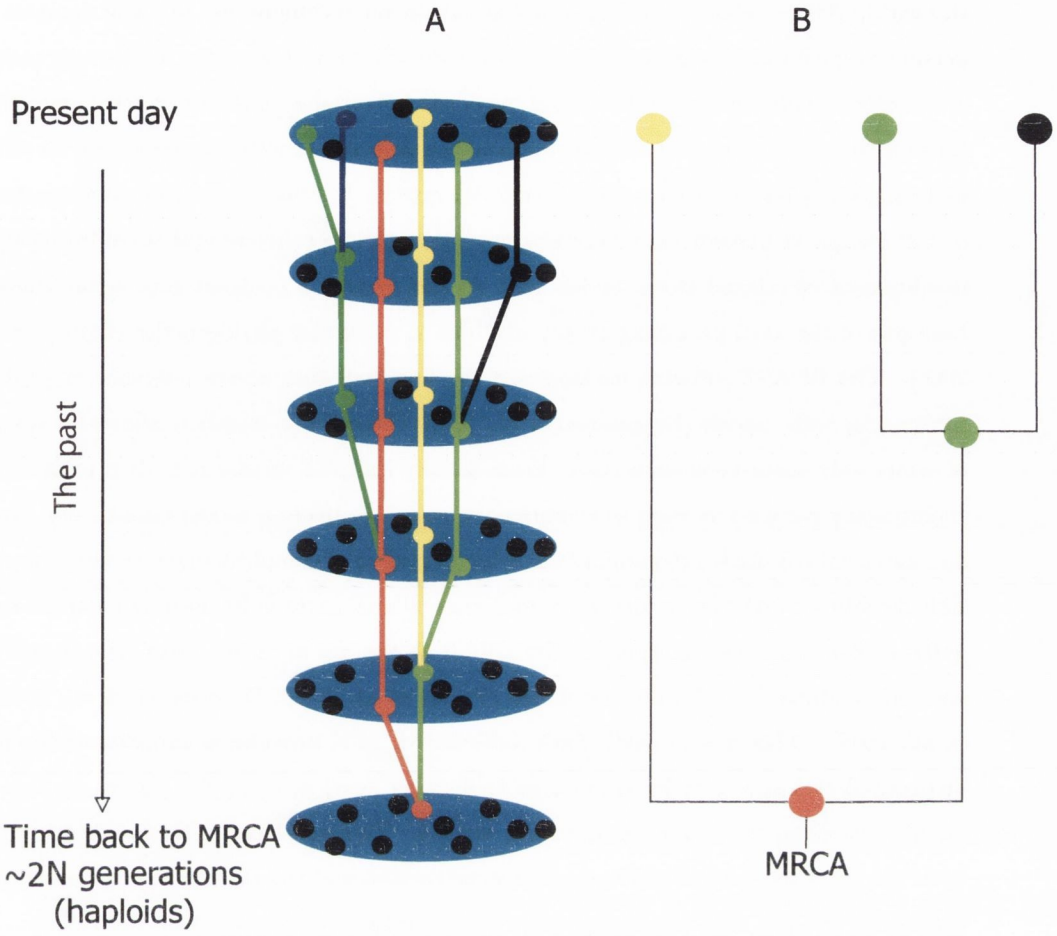


Figure 1.2: The basic principle of coalescent theory. A) Shows a population of haploid individuals where the lines trace the ancestors of the different colour lineages back to a single common ancestor. B) A sub-genealogy for three of the sampled lineages

viruses seem particularly adapted to coalescent based analyses because their rapid evolution generally produces highly resolved phylogenetic trees .Thus, coalescent estimators can be based on a single phylogenetic history. Examples of questions that fall in the realm of coalescent methods include the timing of introduction of a pathogen and how rapid pathogen evolution is within hosts. In fact RNA viruses constitute the most important class of measurably evolving populations (Drummond et al., 2003).

Software packages such as LAMARC (Kuhner et al., 1998; Beerli and Felsenstein, 2001), GENTREE (Griffiths and Tavare, 1994) and BATWING (Wilson et al., 2003) are all useful packages for coalescent based analysis. In particular they are useful for likelihood analysis and are summarised in Stephens (2001). One package of particular interest is the BEAST software package. BEAST is a Bayesian Markov Chain Monte Carlo (MCMC) statistical tool set within a coalescent framework for inferring demographic and population parameters (Drummond and

Rambaut, 2007). MCMC is simply a computational technique for drawing samples from a posterior distribution to get an estimate of the distribution. The power of this approach is that it incorporates phylogenetic information, accounts for variable substitution rates among lineages through the use of relaxed molecular clocks rather than strict molecular clocks and for differences in demographic history from rates of population growth. Furthermore, rates are estimated across a wide sample of plausible trees thereby providing a more coherent statistical framework. The development of relaxed clock models that do not assume a constant rate across lineages has been one of the most promising recent advances in molecular phylogenetics (Drummond et al., 2006). The BEAST software package was the first tool that allows inference of phylogenetic trees under such models (Drummond et al., 2006). More interestingly it allows for the analysis of temporally spaced sequence data. From serially sampled viruses it is then possible to infer evolutionary parameters such as evolutionary rate, substitution model parameters, phylogeny, and ancestral population dynamics from a large number of complementary evolutionary models. The flexible skyline plot is one such tool that allows us to infer past population dynamics without choosing a demographic model apriori as the use of an incorrect demographic model will lead to biased and invalid estimates (Pybus et al., 2000; Drummond et al., 2005; Minin et al., 2008). This tool is particularly informative as it provides a nonparametric change in viral population size, thereby allowing us to discover novel demographic signatures that are not readily described by simple demographic models. This graphical analysis takes into account both the error inherent in phylogenetic reconstruction and the stochastic error intrinsic to the coalescent process, thus producing more correct estimates of statistical uncertainty. Now, it also takes center stage in the evolutionary analysis of RNA viruses where it has provided us with useful insights regarding epidemiology and origin. For example, the application of the Bayesian skyline plot has demonstrated that the Egyptian HCV epidemic was likely caused by viral contamination of injectable antischistosomiasis treatment (Drummond et al., 2005; Pybus et al., 2003b). Further to this BEAST has elucidated that the outbreaks of HIV and HCV occurred before foreign medical staff arrived in Libya (de Oliveira et al., 2006). A full list of publications using BEAST can be seen at `http://beast.bio.ed.ac.uk/Publications`.

Aside from the power of coalescent approaches and the crucial link it plays between phylogenetics and population genetics a number of caveats remain. A strong assumption, and an uncomfortable one, is that selection and recombination are absent and viral sampling is random. The problem of selection can be overcome by first estimating the population structure using genomic data and then testing for selection using the estimated population structure model to construct the null hypothesis. However omitting selection in a coalescent model can also lead to inaccurate estimations of population dynamics. The second problem of recombination makes coalescent based inference exceedingly difficult and adopting suitable models and methods to incorporate such a process have not yet been devised. It is unlikely they will be devised in

the near future (Andrew Rambaut, personal communication). Another problem with coalescent approaches is that these methods are computationally intensive and care must be taken with the choice of prior distributions.

1.3.2 Application of Forensics to RNA viruses

The goal of forensic analysis of pathogens from a microbiological point of view is to establish the pathogen and its source in a scientific based unequivocal manner that can be used as evidence in court. Owing to their rapid rates of evolution RNA viruses can be useful in forensic medicine given the appropriate statistical framework. One of the most well known and scrutinised studies is the "Florida dentist" case (DeBry et al., 1993; Ou et al., 1992; Crandall, 1995; Hillis and Huelsenbeck, 1994; Smith and Waterman, 1992; Holmes et al., 1993a), where by molecular phylogenetic analysis of the *env* region of HIV established support for the transmission of the virus from a HIV positive dentist to six of his patients. Another high profile case following a similar approach using viral phylogenetics revealed that a gastroenterologist injected his former lover with blood or blood products obtained from an HIV infected patient under his care (Metzker et al., 2002). This case was the first time that phylogenetic analysis was used as evidence in a United States criminal proceeding. Phylogenetic analyses proved that patients were not infected by HIV from a Baltimore surgeon (Holmes et al., 1993b). However, caution must be advised when criminal convictions rely solely on phylogenetic evidence coupled with clinical and epidemiological evidence as linkage can be proved but transmission cannot (Pillay et al., 2007). Although phylogenetics provided important insights to a recent investigation on the timing of HIV infections among Libyan children where it was shown that most of the strains of infection were circulating before the accused medical workers arrived thereby vindicating the accused of transmitting HIV (de Oliveira et al., 2006). Similarly, studies have been performed with HCV where the proposals of both the defendant and the prosecutor are translated into phylogenetic hypothesis (Gonzalez-Candelas et al., 2003). These serve as the null and alternative hypothesis from where the likelihood can be determined to transform the probability of two alternative hypotheses. For example in this case whether or not a patient shared the source of infection with other infected patients. In this case the molecular phylogenetic analysis proved that all infected individuals were from a single outbreak in a hospital hemodialysis unit (Gonzalez-Candelas et al., 2003). An improved statistical method was devised and applied to an outbreak of fulminant hepatitis B virus in a public hospital (Bracho et al., 2006b).

In this context molecular epidemiological approaches serve as useful and relevant tools for microbial forensics particularly RNA viruses to support criminal proceedings.

1.3.3 Phylodynamics

In order to understand the transmission of infectious disease it is necessary to unify phylogenetic and epidemiological models. Such melding of the two approaches is called phylodynamics (Grenfell et al., 2004b). A major determinant of epidemic behaviour is the timescale of infection dynamics where the strength of the micro- and macro-evolutionary forces determine the speed of infection. For instance, whether an infection is quickly cleared by the host (acute) or whether it is slower and persists as a chronic infection. Ultimately, the two are influenced by spatio-temporal dynamics and cross protective immunity.

Acute infections are illustrated by those that generate a strong cross immunity and are short in duration. Measles (and other morbilliviruses) demonstrate this end of the phylodynamic spectrum with a strong cross-protective antigenic response. Conversely influenza is an acute respiratory infection with a short infectious period, for which the immune response is only partially cross protective. Lastly, the most common vector-borne viral disease of humans - dengue represents the last category where an immune response may exacerbate disease. This is due to antibody-dependent enhancement (ADE), which induces positive reinforcement between strains (Holmes and Twiddy, 2003). These differences are represented in the shape of their phylogenies (Figure 1.3). The measles virus phylogeny appears complex without obvious divisions due to the co-existence of an immune response that is equally potent against several similar strains. Therefore selection does not operate consistently to leave an imprint (Figure 1.3B). In contrast, the influenza tree is ladder-like reflecting rapid strain turnover as a result of continual immune selection (Figure 1.3A). A dengue virus tree shows serotypes that are phylogenetically equidistant explained from immune enhancement (ADE) that can manifest as dengue haemorrhagic fever and shock syndrome (Figure 1.3C).

The long infectious periods of persistent infections such as HCV and HIV can also be observed in a phylodynamic framework. Phylogenies can be constructed from sequences belonging to different individuals (inter-host) and from sequences evolving within an individual (intra-host). Inter-host dynamics are relatively slow owing to the time between transmission events and thus, relatively slow epidemiological trends are reflected in the phylogenetic structure (similar to dengue and measles) controlled by non-selective population dynamic processes (Figure 1.3B). By contrast, intra-host evolution is driven by continual host immune selection from neutralising antibodies or cytotoxic T lymphocytes with phylogenies showing fast dynamics with limited diversity at any one time (Figure 1.3A).

Grenfell et al. (2004b) constructed a model to accommodate the different phylogenies and temporal and spatial dynamics with the Evolutionary Infectivity Profile (EIP) as a key parameter. The EIP is a measure of the transmission rate of immunologically selected mutations, which reflects the average amount of genetic adaptation transmitted to a susceptible individual. It

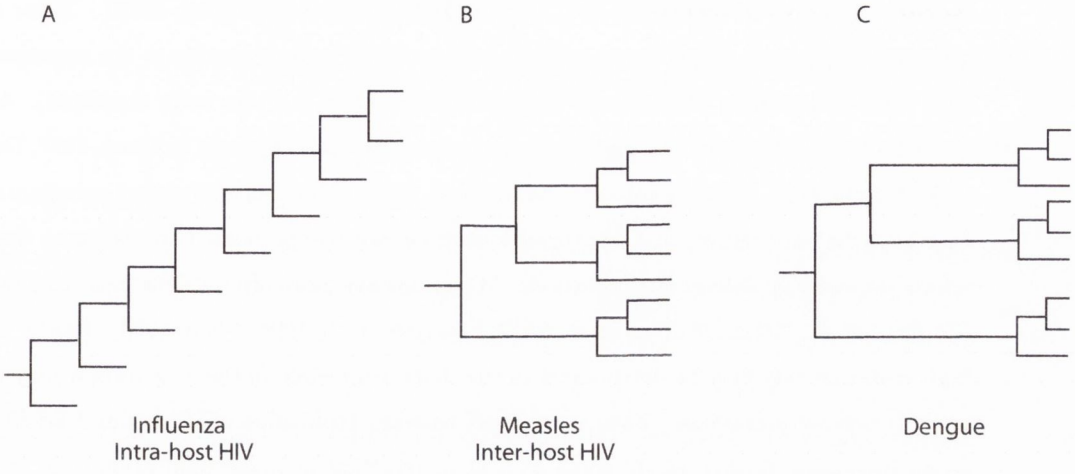


Figure 1.3: Representation of tree shapes under different phylodynamic processes. A) Trees that have a strong temporal structure are illustrated (influenza) and are subject to continual immune driven selection (intra-host HIV). B) and C) Viruses where immune selection is weak or absent are depicted from spatial dynamics or population size as demonstrated by measles and dengue virus.

depicts the changing pattern over time which depending on the immune history of the host may be different. Phylodynamics has already been applied to a number of infectious agents where it captured the evolution of influenza A (H3N2) (Koelle et al., 2006) and revealed the transmission dynamics of the HIV epidemic among men who have sex with men in London (Lewis et al., 2008)

1.4 Rates of evolutionary change

Mutation is the quantum force of genetic variation on which natural selection, genetic drift and recombination all act to shape the architecture of genomes. The mutation rate is the rate at which point mutations, insertions and deletions are generated in a genome. The number of genetic errors can accumulate proportionally to units of time, (per generation or per round of replication). When mutations are neutral there is a relationship between the rate at which they are generated and then fixed at the population level. This refers to the rate of nucleotide substitution, that strictly is defined as the number of fixed mutational changes per nucleotide per unit time (usually measured in years). Duffy et al. (2008) provide an excellent review discussing the patterns and determinants of evolutionary change in viruses.

1.4.1 Mutation Rates

The single most important facet of RNA viruses is their high mutation rate. The range of mutation spans several orders of magnitude from 1.5×10^{-3} mutations per nucleotide, per genomic replication (mut/nt/rep) in the ss RNA phage $Q\beta$, to 1.8×10^{-8} mut/nt/rep in the ds DNA virus

herpes simplex virus type 1 (HSV-1) (Drake, 1993; Drake and Hwang, 2005). These rates of mutation impinge on every aspect of virus biology. RNA viruses mutate at the maximum error rate while maintaining the integrity of genetic information (i.e. the error threshold). Allowing for fast adaptation mediated by advantageous mutations (Domingo and Holland, 1997; Domingo, 2000; Novella et al., 1995; Steinhauer and Holland, 1987). This frequently cited paradigm is solely based on the assumption that beneficial mutations are strong enough to overcome the fitness effects imposed by deleterious mutations. Mutations are more often deleterious than beneficial (Sanjuan et al., 2004b; Malpica et al., 2002; Escarmis et al., 1996; Chao, 1990). Hence, having a high mutation rate may be detrimental in the short term while in the long term it may produce more beneficial mutations. Thus a trade-off between replication efficiency and fidelity exists as an increasing fidelity would come at a cost, resulting in lower replication. An increasing replication fidelity would come at an energetic or kinetic cost and hence have a negative impact on fitness. Although, the cost of replication fidelity has only recently been associated with the evolution of mutation rates in RNA viruses (Furio et al., 2005, 2007). This is discussed later.

1.4.1.1 Measuring mutation rates

Mutation rates are measured in one of two ways: through Luria-Delbruck fluctuation tests or mutation accumulation studies (Drake, 1991). The classical approach to measuring the rates of mutation is the Luria-Delbruck test whereby the frequency of mutations with a certain phenotype arising from clonally replicate expanding populations are measured. The mutation rate is then obtained when the frequency is adjusted to take account of the number of generations and the number of genome replications within each generation. The other commonly used phenotypic measure of estimating mutation rates is derived from mutation accumulation studies, where populations are subjected to bottlenecks and mutational frequency through these changes in population size are noted (Drake, 1991).

1.4.2 Substitution rates

Substitution rates reflect a complex interplay of several factors such as underlying mutation rate, generation time, effective population size and fitness. Therefore, the underlying high mutation rate of RNA viruses compared with DNA organisms means that long-term rates of nucleotide substitution are usually high in RNA viruses. Detailed studies examining the substitution rate of a wide array of RNA viruses have revealed the overall rates of nucleotide substitution fall in the range of 10^{-2} to 10^{-5} nucleotide substitutions per site, per year (subs/site/year) (Jenkins et al., 2002; Hanada et al., 2004). Although most of the viruses exhibit rates in the order of magnitude of 1 x 10^{-3} subs/site/year. This is equivalent to the fixation of 10 substitutions per genome, per year for a RNA virus with a genome length of 10 kb. However even among rapidly evolving

RNA viruses there is variation among substitution rates. This is partially due to the contrasting modes of inter- and intra-host evolution. For instance in HIV-1 the highest rate of evolutionary change is observed at the intra-host level resulting from positive selection that facilitates immune escape (Choisy et al., 2004; Nielsen and Yang, 1998a). A small number of RNA viruses evolve slowly and experience anomalously low rates of nucleotide substitution. For example, Human T-cell lymphotropic virus type II (HTLV-II) varies from 10^{-7} subs/site/year in epidemics, where viruses are maintained within hosts through the clonal expansion of infected cell rather than active replication, to 10^{-4} subs/site/year. Indeed the far higher rates of substitution are observed in HTLV-II populations where transmission and replication are rapid. This is commonly seen in intravenous drug users (Salemi et al., 1999; Vandamme et al., 2000). Another retrovirus showing a reduced rate of evolution is Simian foamy virus (SFV) where values of only 1.7 x 10^{-8} subs/site/year were observed owing to the large latency period within hosts (Switzer et al., 2005).

There have also been a number of suggestions that RNA plant viruses evolve slower than RNA animal viruses. The remarkable genetic stability of plant virus populations is claimed from Tobacco mild green mosaic virus, which has showed no increase in genetic diversity over the 90 years considered (Garcia-Arenal et al., 2001)and Tobamovirus populations which are very stable and do not evolve at a measurable rate (Gibbs et al., 1999). It was even documented that populations of Turnip yellow mosaic virus from Europe and Australia, that probably diverged more than 12,000 years ago, differed by less than 1% (Blok et al., 1987). However, such ideas were recently quashed when, for the first time estimates of evolutionary rates were made using statistical methods developed to analyze temporally spaced sequences from Rice yellow mottle virus (RYMV) (Fargette et al., 2008). Estimates of substitution rate were calculated to be in the order of 10^{-4} subs/site/year and firmly within the range of substitution rates observed in animal RNA viruses (Jenkins et al., 2002). Further evidence from Zucchini yellow mosaic virus (ZYMV), an economically important virus of cucurbit crops support the view that RNA plant viruses like RNA animal viruses are measurably evolving populations (Simmons et al., 2008). While the coat protein genes of Potyviruses have an evolutionary rate of about 1.15×10^{-4} nucleotide subs/site/year (Gibbs et al., 2008). Taken together, these studies confirm and support the view that plant RNA viruses are subjected to the same error prone process of replication and harbour equivalent rates of evolution as animal RNA viruses.

The idea that DNA viruses are usually characterised by far lower rates of substitution than RNA viruses is becoming increasingly less evident with the boundaries between the two groups becoming less clear with respect to rates of evolutionary change. Most notably, studies of two mammalian parvoviruses, canine parvovirus and B19 erythovirus, found substitution rates similar to those of RNA viruses, at $\sim 10^{-4}$ substitutions per site per year (subs/site/year) (Shackelton et al., 2005; Shackelton and Holmes, 2006). In addition the plant geminivirus Tomato yellow

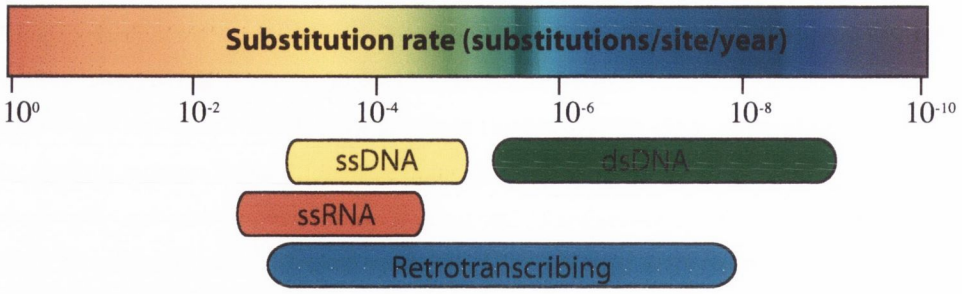


Figure 1.4: The average range of substitution rates for viruses. These ranges are given as substitutions per site per year.

leaf curl virus was observed to evolve at least as rapidly as plant RNA viruses, as quickly as animal ssDNA viruses, and as fast as some animal RNA viruses (Duffy and Holmes, 2008).

1.4.2.1 Measuring substitution rates

Traditionally estimates of substitution rates have been based on phylogenetic methods where the number of nucleotide differences is counted between sequences that are known to have diverged at a specific time point. The most readily used method to estimate viral substitution rates is linear regression. Although useful it suffers from important limitations (Lukashov and Goudsmit, 2002). One such limitation is that there is widespread pseudo-replication because of the pairwise comparisons performed between sequences. Hence deep branches are compared multiple times. Another strong assumption is that of a constant molecular clock where substitutions accumulate at a fixed rate over time. However most RNA viruses now violate this assumption (Jenkins et al., 2002). A recently developed method based on a Bayesian MCMC coalescent framework has attempted to mitigate these problems (Drummond and Rambaut (2007) and section 1.3.1.1)

1.4.3 Causes of high evolutionary rates

Many reasons can be ascribed to why RNA viruses have such a high mutation rate (Belshaw et al., 2008). The underlying differences between mutations rates among taxa can be explained by the trade-off between high-fidelity DNA polymerases with error-correcting mechanisms and low fidelity RNA polymerases, without error correction. In this section I focus on the factors affecting mutation rate, as this also determines the neutral substitution rate.

The mode of life for RNA viruses does not provide sufficient reasoning for their high mutation rate as bacterial microbes attacked by RNA viruses are also high despite the lack of an adaptive immune response (Drake et al., 1998; Drake and Holland, 1999; Drake, 1993). Another tempting explanation, but most likely incorrect, is that the high mutation rate could be an evolutionary constraint for RNA viruses (Drake et al., 1998; Belshaw et al., 2008). Therefore the high error rates of polymerase may simply represent a case that RNA viruses cannot improve on. This is

consistent with genetic information being more compressed in RNA viruses than it is in DNA viruses due to replication and transcription being biochemically equivalent. That it is likely because of the natural variation that exists and the ability of RNA viruses to acquire lower mutation rates (Mansky, 2000; Sanjuan et al., 2004b; Pugachev et al., 2004). RNA viruses are not constrained by evolution and free to reduce their mutation rate in natural populations. Indeed nucleoside analogues such as ribavirin have proved very useful in lowering mutation rates because of the high-fidelity RNA viral polymerases that is created (Crotty et al., 2001; Sierra et al., 2007; Mansky and Bernard, 2000; Pfeiffer and Kirkegaard, 2003).

A more plausible scenario for the high mutation rate of RNA viruses lies in an evolutionary trade-off between replication speed and replication fidelity. The fitness cost associated with replication fidelity may result in viruses been able to replicate quickly or accurately (Elena and Sanjuan, 2005). The cost of fidelity should be stronger for species that rely critically on fast replication, as is the case with RNA viruses whose infection cycles are demonstrative of their parasitic lifestyle (Drake and Holland, 1999). The trade-off is now well established for a number of viruses. Using a series of vesicular stomatitis virus (VSV) mutants carrying single amino acid substitutions in the RNA polymerase gene, Furio et al. (2005) showed that changes leading to lower mutation rates also led to slower growth rates, indicating that fidelity paid a replication speed cost. To shed some light on the biochemical basis of this cost data from previously published in vitro experiments with HIV-1 reverse transcriptase (RT) were analysed (Furio et al., 2007). A positive correlation between the in vitro mutation rate and the catalytic constant for cognate nucleotide incorporation was observed which suggests that an increased fidelity could negatively impact the rate of replication (Furio et al., 2007).

1.4.4 Consequence of evolutionary rates

A significant consequence of high mutation rates is the cap on RNA virus genome size (Holmes, 2003). This expectation is derived from early quasispecies models that the maximum genome sizes is regulated by the reciprocal of their mutation rates (Eigen, 1996b). This error threshold, beyond which selection cannot act, results in the genetic structure being lost and the population randomly drifting in genotypic space. As a result the upper limit of the genome size of a virus is constrained. Alternatively it has been proposed that the limited genome size of RNA viruses may be as a result of biochemical events (independent of mutation rate) arising from the instability of RNA macromolecules (Duffy et al., 2008).

Mutational robustness is defined as a reduced sensitivity to perturbations affecting phenotypic expression. If perturbations are heritable, then we talk about genetic robustness; if they are not (for example, changes in physical and chemical parameters, or developmental noise), then we talk about environmental robustness. Genetic robustness should occur when there are

several copies of a single gene, when several genes contribute to the same function or through biochemical buffering mechanisms (Elena et al., 2006). It is now considered to be a selectable trait (Sanjuan et al., 2007). The mechanism for buffering mutational effects in RNA viruses is still debatable. However it is possible that viruses use chaperones to buffer mutational effects as chaperones are a subset of a highly conserved group of proteins known as heat-shock proteins, that assist in protein folding and aggregation (Hartl et al., 1994; Hartl, 1996). Chaperones are considered to have an important role in maintaining the function of mutated altered proteins by buffering detrimental mutations effects on protein structure stability (Fares et al., 2002b). Upon infection, viruses induce a cellular stress response resulting in the over expression of chaperones (Mayer, 2005). Furthermore, it has been shown that most viruses need cellular chaperones during stages of their infectious cycle where they serve to solve the protein-folding problem and interfere with cellular process (Jockusch et al., 2001). Taken together these observations suggest that such buffering mechanisms are an extrinsic property of RNA viruses generally not encoded in their genomes (Elena et al., 2006).

The consequences of artificially increasing error-rates in RNA viral species have been explored in cell culture experiments. The increase of the mutation rate could actually serve as a therapeutic strategy against RNA viruses. This is the reasoning behind the lethal mutagenesis theory (Bull et al., 2007a; Eigen, 2002). Lethal mutagenesis is a deterministic process that pushes a population to extinction due to elevation of mutation rates to the point where a population is so overwhelmed by deleterious mutations that it cannot maintain itself. This method has been suggested as the basis of successful treatments of viral infections by use of drugs known to elevate mutation rates (Vignuzzi et al., 2005). These drugs commonly referred to as chemical mutagens such as the ribavirin have been used to artificially increase error rates in a variety of RNA viruses, including vesicular stomatitis virus (VSV) (Lee et al., 1997; Holland et al., 1990), human immunodeficiency virus type 1 (HIV-1) (Loeb et al., 1999), poliovirus type 1 (Crotty et al., 2001; Holland et al., 1990), foot-and-mouth disease virus (Sierra et al., 2000), lymphocytic choriomeningitis virus (Grande-Perez et al., 2002), Hantaan virus (Severson et al., 2003), and hepatitis C virus (Zhou et al., 2003). Thus, all these studies provide empirical evidence supporting the principle of lethal mutagensis and proved that chemical mutagens severely reduced viral titers with extinction achieved in some cases.

1.5 Evolutionary mechanisms associated with viral genomes

The rates and consequence of mutations is further confounded by two major evolutionary processes, recombination and positive selection. Both of these mechanisms have a effect on the generation of variability and subsequent fitness and are discussed in detail below.

1.5.1 Recombination

Recombination is considered one of the key evolutionary processes shaping the architecture of genomes. Although nearly all organisms engage in some form of recombination, why recombination occurs and how it is maintained remains controversial (Barton and Charlesworth, 1998; Otto and Lenormand, 2002; Muller, 1964; Hill and Robertson, 1966). In studying the evolution of microbial pathogens an increasing problem is determining the extent to which these genomes recombine. The pursuit of such information is vital because recombination severely complicates the inference of phylogenetic relationships (Posada and Crandall, 2002; Posada, 2000) and similarly confounds reliable estimates of evolutionary parameters such as selection and rates of evolution (Shriner et al., 2003; Schierup and Hein, 2000). Therefore quantifying the effect of recombination is crucial to our understanding of how genetic diversity is generated and maintained in populations.

Recombination involves the rearrangement of genetic material by breaking up linkage between loci and joining previously unassociated DNA fragments. Thus, different regions will have different evolutionary histories. Recombination is usually initiated by a single or double strand break, and is a necessary process for DNA repair and disjunction of chromosomes during meiosis. There are four common types of recombination: homologous recombination, site-specific recombination, transposition, and copy choice or strand transfer. The most common of these is homologous recombination (Lai, 1992). During this process, the donor sequence neatly replaces a homologous region of the acceptor sequence with the recombinant leaving its genomic organisation unchanged. Thus it involves not just homologous parental RNAs, but also crossovers at homologous sites. However this is not always the case as aberrant homologous and non-homologous recombination also occur but at a lesser frequency in the natural environment (Lai, 1992).

There is now a fairly rich literature and reviews documenting recombination in RNA viruses (Posada et al., 2002; Worobey and Holmes, 1999; Awadalla, 2003; Moya et al., 2004). Although most studies indicate that recombination rates in many RNA viruses are often lower than those in other organisms there are notable exceptions (Posada et al., 2002). Perhaps the most dramatic is with retroviruses such as HIV which recombine at high rates through a template switching process during reverse transcription (Jung et al., 2002; Shriner et al., 2004; Dixit and Perelson, 2005; Robertson et al., 1995). In this instance, the presence of recombination between viruses from different primate hosts is thought to be associated with human HIV emergence (Lemey et al., 2006; Keele et al., 2006; Santiago et al., 2005). Recombination also contributes on a wider scale to the global epidemic of HIV with circulating recombinant forms (CRFs) i.e. intra-subtype recombinant transmitted through multiple individuals, playing a role in the epidemiology of the virus (Robertson et al., 2000; Kijak and McCutchan, 2005). However, other than in retroviruses,

recombination is not a particularly common process in RNA viruses. For example, it is a rare or absent mechanism among negative-sense RNA viruses (Chare et al., 2003). This may be a result of RNA packaging as their filamentous ribonucleoprotein is always encapsidated, thereby greatly limiting the template-switching process. Recently it was also concluded that if it occurs at all homologous recombination plays only a very minor role in the evolution of influenza A virus (Boni et al., 2008). In contrast, recombination is a more common phenomenon in positive-sense RNA viruses particularly in plants (Chare and Holmes, 2006; Codoner and Elena, 2008) and some animals (Holmes et al., 1999; Worobey et al., 1999; Heath et al., 2006). Finally, in most cases recombination appears to be a sporadic event that does not occur at a high enough frequency to establish it as a key evolutionary strategy. It could simply be a passive by-product of the replication machinery or ecological circumstances of the virus in question (Holmes and Drummond, 2007).

Frequent recombination seems benefical because it potentially increases fitness by creating advantageous genotypes and purges deleterious mutations from virus populations. However many recombinations are likely to be deleterious as they disrupt optimal protein structures and functional gene combinations. For example the replication proteins of influenza A virus work together as a cohesive complex with alterations of genomic segments reducing the replication efficiency (Taubenberger et al., 2005; Hatta et al., 2002; Clements et al., 1992). Analyses on the functionality of recombinant genes have indicated that the majority of progeny arising from recombination events are non-viable (Meyer et al., 2003; Voigt et al., 2002; Martin et al., 2005b; Escriu et al., 2007; Moreno et al., 2004). This is presumably due to functional constraints with purifying selection acting to maintain proper protein folds (Lefeuvre et al., 2007). Strong selective forces must operate on amino acid interactions that are not preserved for novel recombinants to generate.

1.5.1.1 Methods to detect Recombination

Detecting recombination is far but a simple procedure as it is dependent on a range of factors. During the past 30 years a plethora of methods have been developed to detect and estimate recombination rates. Consequently methods can be broadly divided into parametric and non-parametric methods (Awadalla, 2003).

Parametric approaches measure linkage disequilibrium (e.g. LDhat). The non-parametric approaches fall into main categories: comparative and phylogenetic. Dozens of statistical tests have been developed in the endeavor of identifying recombination (Boni et al., 2007; Kosakovsky Pond et al., 2006a,b; Martin et al., 2005c,a; Martin and Rybicki, 2000; Worobey, 2001; Gibbs et al., 2000; Smith, 1992; Posada and Crandall, 2001; Padidam et al., 1999; Grassly and Holmes, 1997). In addition to testing for the existence of recombination, certain methods are also able to locate

recombination breakpoints and, sometimes, the parent sequences involved in the recombination event, although the latter can be quite difficult. Some approaches search for changes in patterns of genetic diversity (e.g. SimPlot, TOPAL), other methods identify regions that are more similar in base identity, codon usage and base composition (e.g. GENECONV, Max Chi-Square) or look for excessive convergent evolution (Homoplasy Test, PIST). Other methods use phylogenetic tools to identify recombinants or footprints of ancestral recombination events. These methods implicate recombination when regions or genes have different phylogenetic histories (e.g. LARD, PLATO, BOOTSCAN) (Figure 1.5). Many of these methods have now been implemented in a single software package namely the Recombination Detection Package (RDP 3) (Martin et al., 2005c). A full comprehensive list of recombination detection programs is provided at `http://www.bioinf.manchester.ac.uk/recombination/programs.shtml`.

A comprehensive review of statistical methods for detecting recombination is presented in Posada et al. (2002) and references therein. Posada and Crandall (2001) evaluated the ability of 14 non-parametric methods by computer simulations and found that the tests varied greatly in performance depending on the amount of recombination, the genetic diversity of the data and the degree of rate variation across sites. It is clear that reliance on one phylogenetic method may be problematic as many of the phylogenetic methods often give incongruent results (Posada, 2002). In addition other studies have also served useful in the evaluation of methods (Wall, 2000; Wiuf et al., 2001; Brown et al., 2001). Despite all these tests none has yet emerged as a single standard test for identifying recombination. Therefore, as all these authors concluded, one should not rely on a single method to detect recombination but on a combination of various methods. In contrast, other authors suggest that the application of a battery of methods to test for recombination is not helpful (Awadalla, 2003). To understand the role of the force of recombination in generating genetic diversity we need to accurately estimate the rate at which recombination occurs. The current inability of methods to estimate rates of recombination is perhaps the greatest criticism of all approaches.

1.5.2 Natural Selection

The success of Darwin's theory of natural selection is a testament to its ubiquitous nature in shaping morphological and behavioural evolution. However the importance that selection plays in the evolution of genes and genomes is still a matter of debate. This section introduces the basic concepts of positive and negative selection as well as the major theories of molecular evolution.

The neutral theory of molecular evolution claims that most of the observed molecular variation in a population is due to random fixation of selectively neutral mutations or nearly neutral mutations (Kimura, 1968; King and Jukes, 1969). Neutral mutations are those that have little biological fitness significance or are alternatively deleterious. The concept of the neutral the-

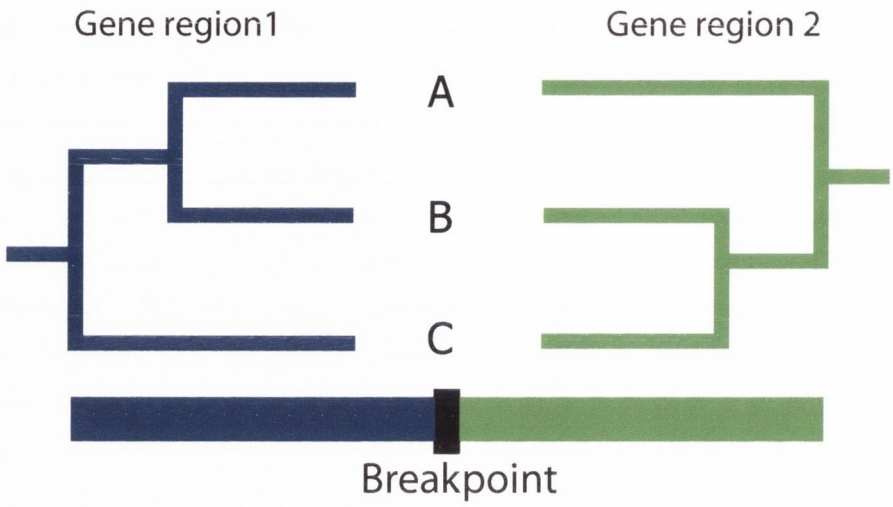


Figure 1.5: One simple test to detect recombination is to search for incongruent trees as different genes produce different trees.

ory of molecular evolution heralded a mini-revolution in evolutionary theory. The controversy surrounding the neutral theory has lead to a number of studies with results leading authors to preclude it as the sole explanation for the evolution of molecules. The neutral theory has helped stimulate interest in the development of analytical tools and has prompted the generation of a number of tests of the theory. See Kreitman (2000); Nielsen (2001); Fay and Wu (2001, 2003); Ohta (2002); Anisimova and Liberles (2007) for reviews.

The nearly neutral theory further expands on prexisting theory by including substitutions that are caused by the random fixation of very slightly deleterious mutations (Ohta, 1973, 1992).

In contrast, the selectionist view states that the majority of genetic variations in an organism are adaptive. Adaptive evolution refers to those amino acid substitutions that confer a certain selective advantage to the organism. Such advantageous mutations are subsequently fixed in the population. The term adaptive evolution is used interchangeably with positive selection throughout this work. Many studies have now concentrated on detecting genes under positive selection as they have become a powerful tool to study disease related genes and thus propose new therapies. The sheer number of adaptively evolving genes ranges from proteins involved in defence or immunity (e.g. viral surface or capsid proteins) to proteins implicated in sexual reproduction (e.g. sperm lysin) (Yokoyama, 2002; Vallender and Lahn, 2004; Yang, 2006). Although not exhaustive this range just gives a flavour of the kind of genes that are under positive selection with a more comprehensive list contained in specialised databases such as the Adaptive Evolution Database (Roth et al., 2005). It is now possible to apply methods for the detection of positive selection at a genome-wide scale. Most scans for positively selected genes have typically been based on the highly similar human, chimpanzee and/or rhesus macaque genomes (Clark

et al., 2003; Nielsen et al., 2005; Bustamante et al., 2005; Arbiza et al., 2006; Bakewell et al., 2007; Gibbs et al., 2007). More recently this has been extended to non-primate mammalian genomes (Kosiol et al., 2008). Together these studies all demonstrate that genes under positive selection are enriched for roles in defense/immunity, sensory perception and reproduction. Clearly there are functional implications for positive selection that have helped to shape present-day genes.

In brief, proponents of the neutral theory of evolution believe that inter and intra polymorphism that we observe today is not due to fixation of advantageous mutation driven by natural selection but to random fixation of neutral mutations. To the contrary, Darwin's theory of natural selection lends to the selectionist theory that most changes in an organism are adaptive.

1.5.2.1 Methods to detect positive selection

The standard method for detecting adaptive molecular evolution in protein coding DNA sequences is by comparing the rates of non-synonymous (nucleotide changes that alter the encoded amino acid) and synonymous (silent nucleotide changes that do not produce an amino acid change) substitutions (Figure 1.6). This ratio is more commonly know as ω or d_N/d_S and represents the most conservative way to measure the intensity of selection (Kimura, 1977; Sharp, 1997; Akashi, 1999; Crandall et al., 1999; Yang and Bielawski, 2000). The symbols K_a and K_s are also commonly used in the literature as alternative notation for $d_N$and d_S specially when these are structured in population rather than species studies . If selection has no effect on fitness, non-synonymous substitution will be fixed at the same rate as synonymous substitution with $\omega = 1$. In the case where non-synonymous mutations are deleterious, the biological fitness of the organism will be reduced and removed by purifying or negative selection. Thus, $\omega < 1$ because non-synonymous replacements will be rarely fixed with $d_S > d_N$. Only when non-synonymous substitutions are selectively favoured they will be fixed at a higher rate than synonymous substitutions resulting in $d_N > d_S$ and hence $\omega > 1$.

Codon based methods are quite powerful at detecting positive selection because they take into account the biochemical parameters that affect the mutational process at the nucleotide level (e.g. transition-transversion ratio). There are now several studies proposing different methods to calculate ω between pairs of taxa (Miyata and Yasunaga, 1980; Perler et al., 1980; Li et al., 1985; Nei and Gojobori, 1986; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995; Moriyama and Powell, 1997; Yang and Nielsen, 2000). However a number of limitations exist with these methods such as the susceptibility to saturation and the suppression of d_S due to RNA level selection possibly leading to the elevation of ω (Resch et al., 2007; Parmley et al., 2006; Mayrose et al., 2007; Chamary et al., 2006).

Firstly ad hoc methods have been proposed using probabilistic models that average a single ω for each gene throughout all lineages and sites (Goldman and Yang, 1994; Pedersen et al.,

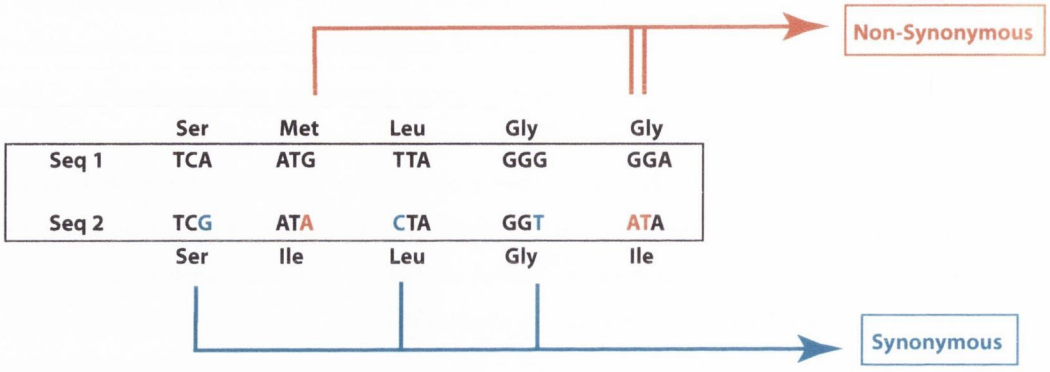


Figure 1.6: Types of mutations in protein coding DNA. These mutations can be classified into synonymous substitutions where nucleotide changes produce no modification at the amino acid sequence level and non-synonymous substitutions where nucleotide substitutions produce amino acid replacements. The three letter amino acid code is used.

1998; Muse and Gaut, 1994). These simplistic models only detect positive selection if it has been strong and persistently evolving (i.e. affects the whole gene or localised to a particular region).

Secondly, these models were further extended to account for the transition/transversion rate bias and permit rate heterogeneity by allowing the ω parameter to vary across sites or regions (Nielsen and Yang, 1998b). This maximum-likelihood approach uses thirteen evolutionary models or statistical distributions from which an ω parameter is drawn to account for the possible occurrences of positive selection (Yang et al., 2000). By performing a likelihood ratio test (LRT) two models can be compared to investigate whether the addition of a model allowing for the presence of positive selection is a better representation of the data than a model that does not. Only if this is the case, is there solid evidence for the presence of positive selection. The good performance of these site-specific models has been well documented (Anisimova et al., 2003; Perez-Losada et al., 2005). Despite the widespread popularity of the single-site method it suffers from a number of criticisms. In fact Suzuki and Nei (2001) claim that it gives spurious results suggesting the use of these methods falsely detect positive selection when there is none.

Moreover the previous models are still conservative as they don't search for punctual diversifying selection across lineages in a tree. This is because the signal for positive selection may be swamped by predominant purifying selection due to a global ω averaged across sites or branches. The extension of the current repertory of codon-based models allows each branch to have its own ω ratio in order to identify codon sites under selective constraints at individual sites in a pre-specified lineage (Yang and Nielsen, 2002). The lineage or subset of lineages is labelled *a priori* and described as the foreground lineage. Remaining lineages are described as the background lineages and are not allowed to have positive selection. A posterior study examining the accuracy of this method found from computer simulations that the model is prone to generate

false positives (Zhang, 2004). However Zhang et al. (2005) later improved the branch site model to give more robust estimates of the ω ratio at selected branches on the phylogeny. Guindon et al. (2004) recently introduced two models that allow selection to switch at sites and branches simultaneously without allowing the approach to become over-parametrised.

An alternative approach to determine site and branch specific ω values without the computational burden of maximum-likelihood is to use counting methods. These methods reconstruct ancestral states using maximum parsimony (Suzuki et al., 2001; Fares et al., 2002a; Berglund et al., 2005) or are distance-based . Methods based on ancestral reconstruction might not provide reliable statistical tests as they ignore errors and biases in reconstructed ancestral sequences. Although the site-class models also suffer with this problem (Yang and Bielawski, 2000). The sliding window based method that optimises the size of the region under investigation in order to obtain the maximum possible resolution has proved to be very useful as it tests for other constraints including accelerated rates of evolution, saturation of synonymous sites, hot spots of substitution and purifying selection (Fares et al., 2002a). Similarly the effect of structural proximity between residues by sliding a three dimensional window instead of a linear window across a protein has been evaluated (Berglund et al., 2005; Suzuki, 2004).

1.6 Epistasis and coevolution

Genes usually interact in a biological system where processes shape the evolution of different inter-specific molecules. Evolutionary interactions between nucleotide sites are thought to be common especially in infectious agents. The processes that shape the evolution of different dependencies are know as epistasis and coevolution.

1.6.1 Epistasis

Epistasis is defined as the deviation from independent gene action on a encoded phenotype. It can be positive or negative depending on the direction of deviation from mutational effects. Negative epistasis between deleterious mutations (also known as synergistic epistasis) may explain the origin and maintenance of sexual reproduction (Kondrashov, 1988). Alternatively, positive epistasis (also know as antagonistic epistasis) maybe a property of compact genomes with few nonpleiotropic biological functions. In fact epistasis correlates with genomic complexity (Sanjuan and Elena, 2006).

Despite its importance little is know about the nature, frequency and intensity of epistasis in natural populations. Viruses, however, may shed light onto the epistasis problem owing to the compact genomes with overlapping reading frames and functional secondary structure. Consequently, in recent years, a number of studies have converged to merge a common picture

in which interactions among deleterious mutations exhibited antagonistic epistasis (Burch and Chao, 2004; Sanjuan et al., 2004a, 2006; Bonhoeffer et al., 2004). This is now considered a rule of RNA genomes owing to their compact genomes with frequent overlap and also results in low mutational robustness. Compensatory mutations may also be caused by interactions between selection and structure. For example, if one substitution leads to a possible improvement in biological function, a second mutation in a coevolving site can be fixed through adaptive evolution. Thereby compensatory changes may salvage deleterious mutations to restore overall fitness making them neutral. One such example of where compensatory changes could occur is in viral epitopes.

A recent highly conservative phylogenetic test for positive epistasis revealed that positive epistasis is commonplace at the nucleotide level in natural populations of RNA viruses and may be central to epidemiological processes such as immune escape (Shapiro et al., 2006). The precise role and evolutionary effects still remain to be elucidated.

1.6.2 Coevolution

The intrinsic complexity surrounding the non-independence of molecules has lead to the emergence of a number of methods focusing on identifying the evolutionary dependencies between amino acid sites. But the coevolutionary relationship between amino acid sites is swamped in a background of stochastic amino acid covariation caused by the non-independence between sites. Furthermore coevolutionary relationships are due to several different factors between two amino acid sites such as functional, structural and physical interactions. Intuitively, each of these factors has different weights and disentangling the different types of coevolution is seldom possible (Fares, 2006). Codoner and Fares (2008) have recently reviewed methods to detect molecular coevolution. Some researchers have used constructed phylogenies to detect coevolution between proteins (Fryxell, 1996; Goh et al., 2000; Ramani and Marcotte, 2003; Pazos et al., 2005; Kim and Subramaniam, 2006; Zheng et al., 2002), others have used mutual information content to determine whether two molecular regions or sites covariate (Codoner et al., 2008; Korber et al., 1993). Alternatively parametric methods have incorporated several correction measures to account for the background noise caused. These are regarded as presenting more statistical power than non-parametric approaches (Pollock et al., 1999; Dimmic et al., 2005; Fares and Travers, 2006; Pei et al., 2006). For example, Fares and Travers (2006) proposed a method capable of distinguishing between background and true coevolutionary occurrences using an amino acid similarity scoring matrix. This method and others have used coevolution as a proxy to detect functional and structural dependencies between residues (Travers and Fares, 2007; Codoner et al., 2006b; McNally and Fares, 2007; Ruano-Rubio and Fares, 2007)

1.7 Debate surrounding RNA virus evolution

The emergence of a new discipline in to the study of RNA virus evolution has brought with it much active debate and controversy surrounding some of the most fundamental questions of all. Much of this controversy is regarding the concept of quasispecies originally proposed by Eigen (Eigen, 1971; Eigen and Schuster, 1977) to model early evolutionary processes and frequently referred to in the viral literature where it is cited whenever populations contain significant genetic variation. Furthermore the so-called survival of the flattest phenomenon and its support for the theory are evaluated and discussed.

1.7.1 The Model of quasispecies

The theory of quasispecies stems from the seminal paper by Eigen in 1971 (Eigen, 1971) with the goal of understanding the origin of life. Later formulated as a mathematical model to refer to an equilibrium mutation-selection process which generates a population of variable genomes (Eigen and Schuster, 1977). This heterogeneous distribution of genomes is organised around one or a degenerate set of fittest sequences know as master sequences (Eigen, 1987, 1993, 1996a; Nowak, 1992). The master sequence will continually generate mutant genomes upon replication where in the absence of genetic drift it will maintain a stable frequency through population time. This is in stark contrast to population genetic models in which genetic drift would occur at neutrally evolving genomic sites. The random sampling of genetic drift is neglected because the concept ignores the effects of population size for organisms with small genomes, large population sizes and high mutation rates. This means that the sequence space surrounding the master sequence may be completely explored, thereby preventing drift from occurring in the population. Another critical element of quasispecies theory is that genomes are not independent entities but are linked by mutational couplings and the entire mutant population evolves as a single unit due to a cooperative structure. In fact this is how the name quasi (acts like) species (single unit) was termed. The formation of this population structure means that natural selection will not act separately on each individual genome but instead on the whole mutant distribution thereby maximising the replication rate of the entire distribution as opposed to a single fittest variant in the population. This strong assumption clearly goes against classical population genetics. An excellent and short review of the topic and its relationship to population genetics theory is provided by Wilke (2005).

The quasispecies was conceived to describe the biochemical kinetics of RNA sequences and as such it seemed natural to apply it to RNA viruses. It has since become a dominant paradigm for RNA virus evolution. Such is the success of this term it is constantly abused by virologists who loosely attach it to sequence polymorphism encountered within viral populations as opposed to its precise foundation. Remarkably the continuous abuse of this term among biologists and

the clear misunderstanding of the implications of the quasispecies theory has never been more evident than where two DNA based organisms with considerably higher genomes sizes and lower mutation rates were implicated with the term "quasispecies" (Blaser, 1997; Gutierrez et al., 1998).

1.7.2 The relevance of quasispecies to RNA viruses

Domingo et al. (1978) were the first to suggest quasispecies as viable model of evolution for RNA viruses. They showed that laboratory populations of $Q\beta$ phage were at an equilibrium distribution of closely related mutants with the consensus sequence stable over multiple passages. Similarly experiments on vesicular stomatitis virus (VSV) showed extreme heterogeneity within clonal populations and a stable consensus sequence maintained over multiple passages in cell culture. However, this view of evolution for VSV was discounted when Jenkins et al. (2001) performed simulation studies using gene sequence data of VSV. The results of these showed that many synonymous sites are neutrally evolving thus creating a sequence space that exceeds the population size of the virus. Variants of the virus could not fully explore all the sequence space surrounding the master sequence and therefore would be subjected to genetic drift, preventing the formation of a quasispecies. This in turn means that natural selection would not act on the entire mutant distribution. It was also highlighted by Jenkins et al. (2001) that a stable consensus sequence is not sufficient evidence to conclude that a population is at equilibrium, the hallmark of a quasispecies.

Other works have also questioned the relevance of quasispecies theory in virus evolution with opponents claiming that quasispecies is a misleading description of RNA virus evolution and pleading for more formal evidence (Holmes and Moya, 2002; Jenkins et al., 2001). Meanwhile proponents of the theory firmly believe in the evidence for quasispecies dynamics in RNA virus populations citing cell culture and *in vivo* work (Domingo et al., 2005, 2006; Domingo, 2002). More recently a study has claimed direct evidence from the analysis of viruses isolated from brain tissue s for a fundamental prediction of the quasispecies theory. A compelling link was proposed between mutation rate, population dynamics and pathogenesis (Vignuzzi et al., 2006). Further discussion of this is found in section 2.5.5.

Wilke (2005) invokes that quasispecies theory is the mathematical equivalent to the theory of mutation-selection balance. A concept that remains one of the most fundamental pillars of population genetics is that natural selection acts on variants increasing their fitness while mutations introduces unfit variants giving rise to a balanced equilibrium distribution between the two effects. The parallels between the two depicts that the real difference between them is that they have been developed largely independently by two separate schools of thought. Despite such stipulation quasispecies theory has its short-comings that need to be addressed but as a

conceptual model it is nevertheless useful.

1.7.3 Survival of the flattest

Darwin's survival of the fittest paradigm suggests that natural selection favours the best-adapted replicators. However the replication rate may not suffice in predicting the eventual survivor because at high mutation rates, the fittest organisms are not necessarily the fastest replicators. It is plausible that the fastest replicators are those that show the greatest robustness against deleterious mutational effects, regardless of the replication costs. This is dubbed the survival of the flattest. In other words, a slowly replicating organism can displace a fast replicating organism if it has strong support from its mutational neighbours to buffer the increasing load (Figure 1.7). This tantalising hypothesis has recently been proven from empirical evidence using digital organisms (Wilke et al., 2001), simulated RNA (Wilke, 2001) subviral RNA plant pathogens (viroids) (Codoner et al., 2006a), and an animal RNA virus (Sanjuan et al., 2007). Codoner et al. (2006a) show that two viroids, which are small single stranded RNAs that do not encode for any protein but yet able to replicate in susceptible plants, were allowed to compete under normal and mutagenic conditions. At a low mutation rate the faster genetically homogeneous species out-competed the slower but highly polymorphic viroid species, as expected from the survival of the fittest effect. In contrast, however when the mutation rate was increased the slower replicator won the competition due to its larger mutational robustness. The manner in which mutational robustness is achieved is largely unknown as it is dependent on the many aspects of the system in question. Although the simplest situation in which robustness evolves at a population level would be when selection is acting across the entire mutant distribution, as opposed to favoring a genotype with the highest replication rate. Such a situation is represented by the quasispecies model of molecular evolution. In another experiment Sanjuan et al. (2007) indicated, using two populations of VSV that the faster replicator outgrew its robust counterpart. However above a critical concentration of chemical mutagens, the competition results were reversed and the flattest more robust VSV population proved the best competitor. Moreover the role that physical space plays in supporting the survival of the flattest theory has been examined under theoretical conditions (Sardanyes et al., 2008).

Indeed these two recent experiments add strong support to the validity of the quasispecies model for viral populations and cast doubts on detractors' arguments that evidence was weak and the quasispecies effect was never supported from *in vivo* work (Jenkins et al., 2001; Holmes and Moya, 2002). Its existence in other biological entities remains to be discovered.

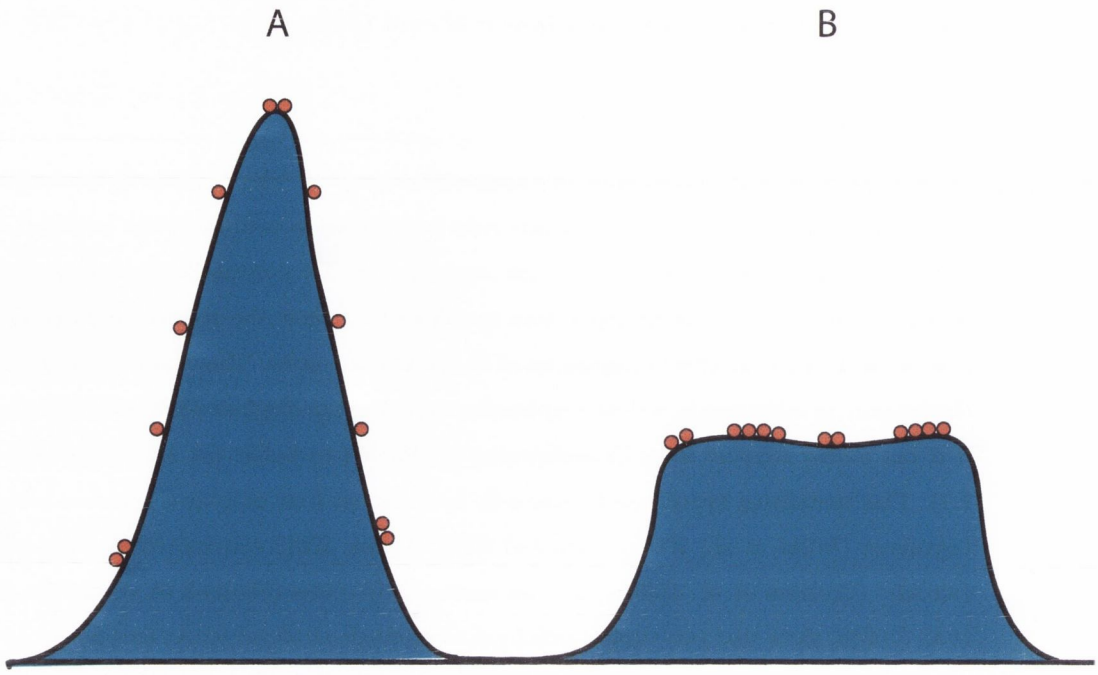


Figure 1.7: Schematic drawing of the survival of the flattest effect. At a high mutation rate, most individuals on the peak A are located at low fitness values, while the individuals on the flat peak B remain close to the local optimum. Therefore the mean fitness of the individuals on peak B exceeds that of the individuals on peak A, an effect which has been termed survival of the flattest.

1.8 Summary, goals and achievements

The battles that exist between humans and their pathogenic viruses are some of the most dramatic ever witnessed for higher organisms. Their stealth of evolution makes them nature's swiftest evolvers and worthy adversaries. The complexities to which RNA viruses evolve have been discussed above and it is these that formulate the theory behind the research chapters. The common goal of all the analyses presented in this thesis is to employ the wealth of genomic data thats available for RNA viruses and to, thus, help elaborate various facets of evolution. The thesis is divided into three main parts each analysing a different RNA virus. The choice of RNA virus to perform such detailed evolutionary-based analyses on was not difficult. First, I selected three major pathogenic viruses that play a dominant role in public health and are of enormous clinical importance. In this regard, the choices of HIV and HCV are not surprising. FMDV, on the other hand although a virus that infects animals it is undoubtedly the single most important veterinary pathogen of our time. Therefore such a choice allows us to perform a direct comparison between RNA viruses that infect solely non-human mammals and those that infect human hosts. Here, I will synthesise those general aspects most relevant to these pathogens.

1.8.1 Human Immunodeficiency Virus (HIV)

There are currently over 33 million people worldwide infected with HIV-1, the causative agent of AIDS (UNAIDS and WHO, 2007). This retrovirus has rapidly disseminated throughout the world to create a pandemic that still poses grave difficulties in producing effective anti viral therapeutics. This difficulty is confounded by the viruses propensity to rapidly evolve creating a huge amount of diversity. It is this complex nature of diversity that challenges our ability to control the virus, hindering vaccine design and development. The enormous genetic diversity of HIV-1 is evident in the classification into three distinct groups (M, N and O), indicating that they arose from three separate transmission events involving SIV from apes to humans (Gao et al., 1999; Sharp et al., 2001; Van Heuverswyn et al., 2006). Group M exists as nine distinct and strongly supported phylogenetic clades labelled subtypes A to D, F to H, J and K all of which have been diversifying since the early 1930s (Korber et al., 2000) in the Democratic Republic of Congo (DRC). This is the most probable location of the origin of the virus (Vidal et al., 2000). Recent phylogenetic evidence derived from a historical sequence obtained in 1960 from an adult female in Kinshas, DRC demonstrates that diversification of HIV-1 in western Africa occured long before the recognised AIDS pandemic (Worobey et al., 2008a). The geographic distribution of this virus is highly heterogeneous representing exportation from the epicentre in central Africa to local epidemics in previously uninfected regions. In order to give you an insight into the revelance of subtypes, I have concentrated on those most important in terms of discovery or global prevalence.

HIV-1 Group M subtype B was the first HIV to be discovered in U.S. populations in the early 1980s and is still the predominant HIV variant in most countries accounting for approximately 12.3% of all infections. Recent evidence has resolved the much debated question regarding the origin and emergence of subtype B and this in turn has generated considerable commentary (Holmes, 2007; Pape et al., 2008; Worobey et al., 2008b). This study elegantly illustrated that this virus travelled out of Africa to Haiti in or around 1966, where it was resident for a number of years before its subsequent migration to mainland United States in or around 1969 (Gilbert et al., 2007). Subtype C is the most well described subtype measured by available full-length genomes as a result of its dominance in the pandemic responsible for >50% of all infections worldwide. It is not clear whether the predominance of subtype C is a reflection of founder events where it has displaced existing HIV-1 subtypes due to different transmission routes or increased sexual activity among host populations (Walker et al., 2005). The other plausible argument for its global dominance may be related to attenuation of virulence. Early findings lead us to believe that subtype C may be less virulent compared with other subtypes, resulting in slower disease progression and longer periods of asymptomatic infection allowing more opportunities for transmission and spread (Arien et al., 2007). However a number of studies point to the

suggestion that subtype C is transmitted efficiently as other group M subtypes (Walker et al., 2005; Ndung'u et al., 2001; Renjifo et al., 1999; Ball et al., 2003). Furthermore there is increasing evidence for the more popular assumption that subtype C possesses a relatively higher fitness level and is more virulent compared to other subtypes (Neilson et al., 1999; Gray et al., 2005). While subtype C is concentrated in Southern and East Africa and in India its prevalence has rapidly expanded in Brazil and similarly in China.

1.8.2 Hepatitis C Virus (HCV)

HCV is the most widely studied and best understood positive-sense RNA virus, largely because it is the leading etiologic agent of chronic liver disease. An estimated 170 million people worldwide are at risk of liver disease with 70-85% caused by HCV developing chronic infections (WHO, 1999). While the virus is responsible for approximately 10,000 deaths per year in the U.S. alone. Despite the absence of a satisfactory animal model, HCV is a excellent candidate system for studying genetic diversity owing to the increasing wealth of sequence data being generated. The diversity of HCV is reflected in the existence of six mutually exclusive virus types termed genotypes (Robertson et al., 1998; Simmonds et al., 1993, 2005). Each genotype is phylogenetically subdivided into a number of subtypes labelled alphabetically in their order of discovery. These types and subtypes exhibit complex patterns of geographic distribution, prevalence and modes of transmission. They have been loosely attached to three groups. The first so called 'epidemic' group contains subtypes 1a, 1b, 2a, 2b and 3a which have a global distribution arising from their route of transmission, namely infected blood products and intravenous drug users (Simmonds, 2004). The next group of HCV strains is termed the 'endemic' group as they are less prevalent and are geographically restricted. Subtypes belonging to type 6 match this description as infections are confined to South East Asia and it is generally thought that the high diversity represents a much older origin for these strains in comparison to the rest. Lastly, a 'local epidemic' group of strains containing subtypes such as 4a are found at increasing prevalence but only at specific locations and among particular risk groups. For example, in Egypt over 10% of infections are attributed to subtype 4a while infection outside the Middle East is rare. Taken together transmission routes play a pivotal role in accounting for the varied epidemiological behaviour although genetic differences between strains has never been fully subjected to detailed study.

1.8.3 Foot-and-Mouth Disease Virus (FMDV)

FMDV is the causative agent of an acute, systemic vesicular disease that affects domesticated and wild cloven-hoofed animals. It is a small non-enveloped virus with a positive sense single-stranded RNA genome that belongs to the *Apthovirus* genus in the family *Picornaviridae*.

FMDV is classified into seven immunologically diverse serotypes distributed throughout the world (South Africa Territories (SAT) 1-3, Asia-1, A, O and C) reflecting its huge spectrum for genetic and antigenic diversity (Knowles and Samuel, 2003). Serotype O is the most predominant of the serotypes and was implicated in the past outbreaks in the United Kingdom where it caused serious devastation resulting in huge economic loss. Type A contains a greater geographical spread than any of the other serotypes. It is also the most antigenically diverse of the Eurasian serotypes with a large amount of variants in Asia, Africa and South America (Ansell et al., 1994; Konig et al., 2001; Araujo et al., 2002; Tosh et al., 2002). To date according to the OIE/FAO World Reference Laboratory for FMD, the last type C outbreak to emerge was in East Africa and Asia in 1996. The virus type Asia1 is endemic to southern Asia and is now considered the least diverse type with a single subtype (Ansell et al., 1994). South African territories FMDV types (SAT 1-3) have a natural reservoir in the African buffalo (*Syncerus caffer*) (Condy et al., 1985). They are also characterised by their large degree of antigenic diversity (Bastos et al., 2003; Sangare et al., 2003). Foot and mouth disease is considered to be the most economically devastating livestock disease in the world, and represents a worst-case scenario for livestock due to the wide variety of species at risk, its rapid spread and difficulty in controlling outbreaks. The 2001 outbreak in Great Britain resulted in the slaughter of more than 6 million animals and an estimated economic loss of 20 billion dollars. The World Organization for Animal Health (OIE) classifies FMD as a listed disease, which requires immediate notification (within 24 hours) due its rapid spread and impact on international trade of animals and animal products.

1.8.4 Achievements

This thesis can be broadly divided into two parts. The first part of this thesis, Chapters 2-4, is solely based on the animal virus FMDV. In Chapter 2, I use a single gene, VP1, that contains the major antigenic determinants for the virus and is essential in cell receptor recognition and escape from the host immune response. The application of a variety of phylogenetic and evolutionary methods allows us to understand the evolutionary driving mechanisms behind FMDV serotypes, sequence diversity, spread and divergence of viral isolates. This single gene study indicates that coevolution of amino acid sites from antigenic epitopes under positive selection pinpoints the functional communication that occurs within molecules. The location of amino acid replacements in different groups was also of interest as heterogeneity of selection forces is operating across serotypes. In the following chapter, Chapter 3, to fully understand the genetic variation of evolutionary rates in FMDV a comprehensive survey of nucleotide substitution rates was carried out using new relaxed clock models incorporated into a Bayesian MCMC approach. In addition a time-line of diversification for the disease is established for the very first time with the initial

radiation of FMDV coinciding with the dawn of European exploration. I discuss the ways in which European exploration may have triggered the prehistoric emergence of FMDV and fostered their speciation. In Chapter 4 comparative analyses of complete FMDV genomes between all seven serotypes are presented. A dichotomy between structurally exposed capsid proteins and non-structural proteins is characterised owing to the greater number of adaptive mutations and relaxed selection in structural proteins coupled with compensatory mutational dynamics from protein structural data. More importantly the pathway of evolution for those serotypes responsible for epizootics on the continent of Africa is found with shifts in the selection-drift balance largely responsible.

In the second part of this thesis, Chapters 5 and 6, I examine two of the most serious human RNA viruses. In Chapter 5 I explore the functional and interaction prediction that coevolution can make and apply this to the pandemic HIV-1/AIDS group. The use of coevolution as a proxy allows us to detect functional dependencies in the *env* gene where the extent of coevolution identified between amino acid residues reflects the functional co-dependence of the gp120-gp41 trimer. Thus, coevolution reflects the complexity of evolution operating within *env*. The insights gained from coevolution analysis also proves an accurate model for the adaptation of HIV-1 to humans as amino acid residues seldom evolve totally independently from one another. Moreover, intra-molecular sites should be subject to coevolution quite often since they have to maintain their physical chemical characteristics in order to ensure a stabilised interaction.

Finally, in Chapter 6 I provide the first evidence for a model of positive Darwinian selection in ancestral lineages leading to genotypes belonging to HCV. The analyses presented reveal a pattern of adaptive mutations becoming fixed in structural genes while non-structural genes are enriched by purifying selection. Furthermore it appears that there is no common underlying mechanism of evolution operating on strains with genetic drift accounting for a large proportion of the sequence diversity observed between types

Collectively these research chapters demonstrate the complexities encountered in unravelling the evolution of RNA viruses as infectious agents of animals and humans. Such complexity can be highlighted by the amount of genetic diversity each virus contains. Therefore, determining the evolutionary and population genetic parameters of RNA viruses is an essential step to minimise their economic and social devastation effects. The different evolutionary strategies unravelled in this study reflect the plasticity of RNA viruses to control and overcome their demographic and epidemiological histories.

Chapter 2

Unravelling selection shifts among Foot-and-Mouth Disease Virus serotypes

The research decribed in this chapter has been published in *Evolutionary Bioinformatics online* (Tully and Fares, 2006).

2.1 Abstract

FMDV virus has been increasingly recognised as the most economically severe animal virus with a remarkable degree of antigenic diversity. Using an integrative evolutionary and computational approach we have compelling evidence for heterogeneity in the selection forces shaping the evolution of the seven different FMDV serotypes. Our results show that positive Darwinian selection has governed the evolution of the major antigenic regions of serotypes A, Asia1, O, SAT1 and SAT2, but not C or SAT3. Co-evolution between sites from antigenic regions under positive selection pinpoints their functional communication to generate immune-escape mutants while maintaining their ability to recognise the host-cell receptors. Neural network and functional divergence analyses strongly point to selection shifts between the different serotypes. Our results suggest that, unlike African FMDV serotypes, serotypes with wide geographical distribution have accumulated compensatory mutations as a strategy to ameliorate the effect of slightly deleterious mutations fixed by genetic drift. This strategy may have provided the virus by a flexibility to generate immune-escape mutants and yet recognise host-cell receptors. African serotypes presented no evidence for compensatory mutations. Our results support heterogeneous selective constraints affecting the different serotypes. This points to the possible accelerated rates of

evolution diverging serotypes sharing geographical locations as to ameliorate the competition for the host.

2.2 Introduction

Foot-and-Mouth Disease Virus (FMDV), a single-stranded RNA virus, belongs to the Picornaviridae family. It has a genome of 8.5 kb that is translated into a polyprotein and processed to yield the structural (P1) and non-structural (P2 and P3) proteins Belsham (1993). P1 produces the four different capsid proteins (VP1 to VP4). VP1 is the most surface-exposed capsid protein (Acharya et al., 1989) and contains the major antigenic determinants A, C and D . Site D is also formed by residues from VP2 and VP3 and it involves residues of the C-terminus of VP1 (Lea et al., 1994). The antigenic diversity of VP1 makes it an ideal candidate for epidemiological studies (Bastos et al., 2000, 2001, 2003; Knowles and Samuel, 2003; Sangare et al., 2003, 2001).

FMDV is classified into seven immunologically diverse serotypes distributed throughout the world (South Africa Territories (SAT) 1-3, Asia-1, A, O and C). It has an enormously wide spectrum of infection reflecting its genetic and antigenic diversity. O is the most predominant of the serotypes. Type A possesses greater geographical spread than any of the other serotypes. It is also the most antigenically diverse of the Eurasian serotypes with a large amount of variants in Asia, Africa and South America (Ansell et al., 1994; Konig et al., 2001; Araujo et al., 2002; Tosh et al., 2002). To date according to the OIE/FAO World Reference Laboratory for FMD, the last type C outbreak to emerge was in East Africa and Asia in 1996. The virus type Asia1 is endemic to southern Asia and is now considered the least diverse type with a single subtype (Ansell et al., 1994). South African territories FMDV types (SAT 1-3) have a natural reservoir in the African buffalo (Syncerus caffer) (Condy et al., 1985). They are also characterised by their large degree of antigenic diversity (Bastos et al., 2003; Sangare et al., 2003) .

A number of evolutionary and population parameters are believed to be responsible for the emergence and dispersal of the different virus serotypes. These constraints can be broadly classed as genetic, for example the ability of the virus to avoid the host immune response, or ecological such as host mobility and population density. Elucidating the mechanisms utilised by viruses that subsequently drive the evolution and infectivity would critically aid in our understanding of viral epidemiology and disease prevention. Recently, Carrillo et al. (2005) have performed an exhaustive comparative genomic analysis to identify novel functional constraints acting on the genome. Conversely, we explored all the available sequences for a gene belonging to the different FMDV serotypes. Analysis of the constraints governing FMDV VP1 protein has been previously conducted to determine the main causes for the emergence of FMDV strain CS (serotype C) (Elena et al., 1992). This insightful work yielded precise information about the evolutionary parameters governing FMDV serotypes evolution. However, heterogeneity in the evolutionary

constraints has not been explored with much detail due to a limitation in the number of sequences and computational tools. Also, no detailed study has been performed before to exhaustively explore other selective constraints at the molecular level in FMDV using an unbiased sample of sequences.

It has been long recognised that the VP1 gene differs in about 30-50% between serotypes (Knowles and Samuel, 2003). We postulate that there are differences in the serotypes' ability to infect and spread that may be accounted for by heterogeneity of evolutionary pressures. In this study we test for changes in the selection constraints to which amino acid sites are subjected by examining variations in the evolutionary rates of amino acid sites (selection shifts or site-rate changes between lineages) between the different FMDV serotypes. For this analysis we use the capsid protein VP1, which is essential in cell receptor recognition and escape from the host immune response (Sobrino et al., 2001). We ask the question of what selective constraints have very likely been responsible for the emergence and spread of the different FMDV serotypes. To test for heterogeneity in selective constraints among FMDV serotypes, we conducted several types of analyses. We first determined whether different selective constraints have shaped the evolution of FMDV serotypes by testing for accelerated fixation rates of amino acid substitutions at particular lineages. As a second approach to the identification of selection shifts between serotypes, we also tested functional divergence (within-site amino acid substitution rates changes throughout the phylogeny) between serotypes. Thirdly, amino acid sites responsible for within-serotype accelerated rates of evolution or functional divergence are identified and their biological importance for the ability of the virus to spread and infect examined. Finally, we identified the co-evolutionary relationships between amino acid sites under accelerated rates of evolution or functional divergence and other protein regions as to understand the molecular consequences of such changes in the VP1 protein of FMDV and to speculate about possible epidemiological scenarios responsible for such selective shifts. We perform an exhaustive analysis of how fixation of amino acid replacements by positive Darwinian selection, coevolution between functionally/structurally important residues and functional divergence in VP1 has performed an efficient strategy for FMDV to spread. This study complements previous reports on the evolution and emergence of foot-and-mouth disease serotypes and pinpoints the need for further experimental work on apparently non-important protein regions that extend beyond well-recognised antigenic sites.

2.3 Materials and Methods

2.3.1 Sequence data

A dataset comprising 665 FMDV VP1 sequences was downloaded from Genebank. This includes all serotypes and only comprises sequences with length greater than 400 nucleotides to account for most of the antigenic regions within the VP1 protein. Table 1 of Supplementary Information contained at `http://www.la-press.com/journal.php?journal_id=17&issue_id=30` describes the serotypes and number of sequences used including the range of sampling times. Accession numbers of the sequences used in this study are provided in Table 2 from `http://www.la-press.com/journal.php?journal_id=17&issue_id=30`. All serotypes are significantly represented, both geographically and sample wise.

2.3.2 Sequence alignments and phylogenetic tree inference

Alignments were obtained for VP1 amino acid sequences for each individual serotype using the T-coffee program (Notredame et al., 2000). We then aligned protein-coding nucleotide sequences for each one of the serotypes by concatenating triplets of nucleotides based on the amino acid alignment. An alignment including the 665 sequences was also built as explained above to test for functional divergence type I (e.g., within-site rates changes among serotypes, see below for a more detailed definition).

To infer an accurate phylogenetic tree, we first estimated the shape parameter (α) of the gamma distribution for the amino acid sequence alignment using the program AAml from the PAML package version 3.14 (Yang, 1997). This parameter takes into account both multiple amino acid substitutions per site and unequal substitution rates among sites. The substitution model WAG, the model that significantly improves the log-likelihood value of the data after comparison with other models available in the program ModelTest (Posada and Crandall, 1998) using the Likelihood Ratio test (LRT), was used to estimate α. Finally, we inferred the phylogeny by neighbour-joining(Saitou and Nei, 1987) using gamma-corrected amino acid distances, estimated by maximum likelihood under the WAG model for the dataset of FMDV sequences, using the program MEGA v3.0 (Kumar et al., 2004)

2.3.3 A first test of changes in selective constraints using neural networks

To test the biological (functional / structural) importance of particular amino acid sites in the VP1 protein we used two approaches. The first approach was based on the collection of the functional information available this protein. However, most of the amino acid sites in the molecule have not been tested for their functional importance since studies have been

biased towards testing antigenic peptide regions. We hence conducted a neural-network analysis implemented in the program CONSEQ (Berezin et al., 2004). We used as input alignment each one of the FMDV subtype alignments (including each one a minimum of 30 sequences and a maximum of 203 sequences (Table 1 at `http://www.la-press.com/journal.php?journal_id=17&issue_id=30`). Briefly, CONSEQ derives automatically a phylogenetic tree from the query multiple sequence alignment and calculates the substitution rates at each position in the sequence alignment by maximum likelihood. A neural network analyses is thereafter conducted by CONSEQ to predict schemes to discriminate between buried and exposed residues. Although the sensitivity regarding the buried and exposed states has been reported to be of about 56% (Berezin et al., 2004), we tested the performance of the test by comparing the predictions to the positions of the amino acid sites in the available crystal structure for the VP1 protein of FMDV. The functional or structural importance of sites was determined by the conservation of the site, with significantly exposed conserved sites being functionally important and conserved buried sites structurally important.

2.3.4 Characterising the main selective constraints in FMDV serotypes

To test if the different FMDV serotypes have been subjected to different evolutionary constraints at the molecular level, we conducted a precise analysis of selective constraints for each serotype. Normally, the intensity of selection on protein-coding sequences is measured by comparing the number of substitutions per synonymous site (d_S) to the number of replacements per non-synonymous site (d_N) (Kimura, 1977; Sharp, 1997; Akashi, 1999; Crandall et al., 1999; Yang and Bielawski, 2000). The ratio between the two rates ($\omega = d_N/d_s$) helps to elucidate if the gene has been fixing amino acid replacements neutrally ($\omega = 1$), replacements have been removed by purifying selection ($\omega < 1$), or mutations have been fixed by adaptive evolution ($\omega > 1$). Furthermore, since d_S values are neutrally fixed (except when constraints operate at the RNA level in these sites) and unless synonymous sites are saturated they are proportional to the time since the sequences compared have diverged. Consequently, the normalisation of d_N values by d_S makes ω independent of time. It has been shown, however, that ω is a poor indicator of the action of adaptive evolution (Sharp, 1997; Crandall et al., 1999). The rationale behind this is that punctual events of positive selection of advantageous non-synonymous mutations throughout the evolution of a particular sequence (episodic positive selection) may be hidden by an overwhelming number of deleterious mutations removed by purifying selection from the same protein regions. Consequently, the number of amino acid substitutions will be below the expectation under neutrality despite these punctual events of adaptive evolution and the mean ω values will be < 1. A way to tackle this problem is by estimating ω value for specific branches of the tree and for each codon region of a protein. Unless branches are pre-specified, using

maximum likelihood for this kind of analysis is computationally prohibitive especially when the number of sequences is large and the model is complex.

We have used Kimura-based models to measure the intensity of selection between and within serotypes using the set of 665 sequences. To detect selective constraints we used a sliding-window based approach (Fares et al., 2002a) implemented in the program SWAPSC version 1.0 (Fares, 2004). Briefly, the program slides a statistically optimum window size along the sequence alignment to detect selective constraints and estimates the probability of d_N and d_S using a number of simulated data sets. We used 1000 simulated data sets in our analysis obtained using the program EVOLVER from the PAML (Yang, 1997) package, taking as initial parameters the average ω, transition-to-transversion rates ratio and codon frequency table generated under the Goldman and Yang model (Goldman and Yang, 1994). These parameters were estimated from the real sequence alignment. The program then slides the window along the real sequence alignment and estimates d_N and d_S by the method of Li (1993). The significance of these estimates is tested assuming a Poisson distribution of nucleotide substitutions along the alignment. This program was also used due to its ability to identify adaptive evolution ($\omega > 1$) but also accelerated rates of amino acid substitutions without the restriction of $\omega > 1$ (that can be due to either adaptive Darwinian evolution or to the fixation of slightly deleterious mutations by genetic drift), saturations of synonymous sites (where d_S values are underestimated) and hot spots (where both d_S and d_N are significantly high but where $\omega < 1$).

2.3.5 Testing Functional divergence between serotypes

Normally, in phylogenetic tree inference and detection of selective constraints, fast evolving sites and slowly evolving sites are considered to remain fast and slow sites, respectively, throughout the entire evolutionary history (i.e. the functional importance of sites has not changed along the evolution of the gene, and thus the rate of evolution of that site remained constant), something called the homogeneous gamma model or rates-across-site (RAS) model (Gaucher et al., 2002). If the functional/structural importance of particular amino acid sites changes between lineages, their rates of evolution also changes and a non-homogenous gamma model will explain better the process of evolution.

A non-homogeneous gamma model of evolution can be statistically described by the functional divergence type I (Gu, 1999, 2001; Wang and Gu, 2001). Because of the significant serotypic and geographic split between FMDV types, we tested whether this split between serotypes was reinforced by serotype-specific amino acid replacement rates. We tested this selection shift by comparing the likelihood value of the hypothesis of functional divergence to that of a RAS model. This comparison was performed using the Likelihood-ratio test (LRT). We tested functional divergence type I between serotypes using the program DIVERGE version

1.04 (Gu, 1999). DIVERGE tests for functional divergence type I by estimating the parameter of functional divergence (θ), which is equivalent to the anti-correlation in the evolution of the two groups of sequences (e.g., FMDV serotypes) being compared. It estimates the posterior Bayesian probability for an amino acid site to belong to the group of sites responsible for functional divergence. These sites share the property of being highly constant in one of the sequence groups and variable in the other, supporting the gain/loss of a functional/structural constraint in one of the groups.

2.3.6 Identifying coevolution between structural/functional regions in VP1

Amongst FMDV antigenic sites in VP1 protein, Site A is involved in conferring the virus its ability to escape from immune pressure and to recognise host cell receptors (Aggarwal and Barnett, 2002). Important structural/functional residues spatially adjacent to these sites may have contributed to the antigenicity of the VP1 protein by their functional or structural interaction. In fact, coevolution between sites surrounding functionally important protein regions allows maintenance of the structural and spatial stability of active sites (Gloor et al., 2005). Fixation of amino acid replacements in regions with strong functional/structural constraints can improve the biological fitness of the virus (for example to escape from the immune response of the host). Alternatively, slightly deleterious mutations (mutations compromising the structural stability or functional performance of the protein) can be fixed under genetic drift in or nearby functional/structural regions and be compensated by conditionally advantageous mutations (for example, mutations that are advantageous due to their compensatory effect) at interacting sites. The mutational dynamic of the protein (for example, fixation of slightly deleterious mutations, advantageous mutations or compensatory mutations) yields information hence on the evolutionary pathways that a group of sequences (serotype) have undergone and provide support for the functional and/or structural importance of a specific molecular region. Analysis of the different types of mutations also provides an opportunity to test the heterogeneity in the mutational dynamics between serotypes.

We tested intra-molecular coevolution in each one of the serotypes by applying a method based on the information theoretic quantity called mutual information to measure the dependence of mutations in VP1 (see Korber et al. (1993) for details). This method measures the covariation between two amino acid sites using the Mutual Information Criterion (MIC). Mutual information is represented by the entropies that involves the joint probability distribution, $P(s_i, s'_j)$, of occurrence of symbol i at position s and j at position s'. The MIC values generated range between 0, indicating independent evolution, and a positive value whose magnitude depends on the amount of covariation. We only included in the analysis those variable positions in the

alignment that were parsimony-informative (i.e. they contain at least two types of amino acids and at least two of them occur with a minimum frequency of two). The statistical power of the test is dependent on the number of sequences as well as the amount of variability at each amino acid site. In this case, both parameters were secured since the number of sequences was high and the variability as well. The significance of the MIC values was assessed by randomisation of pairs of sites in the alignment, calculation of their MIC values and comparison of the real MIC values with the randomly generated distribution of MIC values. One million permutations were conducted to test MIC value significance.

2.4 Results

2.4.1 Heterogeneous patterns of adaptive evolution between serotypes

All serotypes (SAT 1, SAT 2, Asia1, A and O), except C and SAT3, presented evidence of positive Darwinian selection in the VP1 protein (Table 2.1). We have also analysed the functional/structural importance of these sites using a neural-network based method implemented in the program CONSEQ (Berezin et al. 2004). The assumption made in this program is that conserved regions at the amino acid level may have a functional or structural role. We did not consider the buried/exposed state for the amino acid sites inferred by CONSEQ due to the low sensitivity (56%) reported for this method regarding these states. This analysis supported that regions detected as having undergone adaptive evolution present evidence of a functional or structural role (Table 2.1). The comparison of the five serotypes under positive selection highlights diverse patterns of selective constraints at the molecular and phylogenetic levels. Interestingly, all those serotypes under positive selection comprise a similar region around N131- R149 that is located in the immunodominant site A within the G-H loop. However only serotypes A and O, the wider geographically spread FMDV serotypes, have the full RGD motif under positive selection. Moreover, serotypes SAT1 and 2 present evidence of adaptive evolution at overlapping VP1 protein regions. In fact, all of the regions detected in SAT1 to be under adaptive evolution were also found in SAT2 to be positively selected (Table 2.1). Interestingly, SAT2 presented greater percentage of adaptive evolution compared to SAT1, which correlates with a wider geographical distribution of this serotype in Africa. Conversely, regions under positive selection in the widely geographically distributed serotypes (serotypes A, Asia 1 and O) presented poor overlap. Interestingly, there is more overlap in the regions under adaptive evolution between SAT serotypes and serotypes Asia1, O and A than within widely spread serotypes (for example between Asia1 and O). If we were to classify these serotypes in terms of percentage of codon sites under positive selection the following is observed: SAT2 > ASIA 1 > A > SAT 1 > O. The

difference from SAT 2 to type O represents a 10-fold decrease in the percentage of sites under adaptive evolution

Table 2.1: Amino acid sequence regions for the VP1 protein detected as having undergone positive selection in each FMDV serotype using the program SWAPSC. Regions in bold are those detected in SAT1 and SAT2. F ans S refer to functionally and structurally important sites, respectively, as predicted by a neural network analysis. Functionality refers to the antigenicity of those sites. Non-synonymous-to-synonymous rates ratio (w). NN refers to Neural network.

Serotype	Region	Functional	NN Predicition	ω
SAT-2	18-40	Site 3	F, S	3.869
	43-62	BC loop/Site 3	F, S	4.117
	65-68		F, S	2.207
	74-77		F	8.754
	81-87		F	3.525
	97-101		F	4.058
	108-114		F, S	2.903
	117-120		F, S	1.813
	123-128		F, S	4.015
	134-143	Site A	F, S	3.000
	151-162	G-H Loop	F, S	4.634
	172-178		F, S	2.886
	189-192	Site D/C	F, S	8.111
	196-203	C-T loop/Site C	F	6.399
	213-218		F	5.067
SAT-1	45-50	B-C loop	F	1.526
	95-101		F	1.658
	136-149	G-H loop	F	7.080
	199-211	Site D/C	F	7.662
	216-244		F	5.313
ASIA 1	92-105		F	6.540
	135-143	G-H loop	F	4.982
	148-160	Site A	F, S	3.356
	209-214		F	8.965

Serotype	Region	Functional	NN Predicition	ω
A	44-51	B-C loop/Site 3	S	2.030
	131-148	G-H loop	F	1.415
	164-217	H-I loop/Site D/ Site C	F, S	2.586
O	4-6		F	5.311
	68-70		F	2.500
	81-83		F	2.167
	85-87		S	3.107
	96-98			2.500
	137-142	G-H loop		3.371
	144-153	G-H loop	F	3.216
	177-179		F, S	2.682
	197-199	C-T loop/Site C		8.131
	201-204	C-T loop/Site C	F	2.167
	211-231		F	4.782

2.4.2 Evidence for changes in selection constraints among FMDV serotypes

For each individual serotype we applied the neural-network analysis to detect functional/structural sites implemented in the program CONSEQ. This analysis also enabled us to know whether sites under adaptive evolution had any functional/structural role. In the seven serotypes, the RGD motif from the VP1 protein was identified as functionally important, except in the case of Asia1 and A serotypes where only the G144 and D145 were identified as important. The level of functional important sites varied significantly between serotypes (Figure 2.1). The variability of the proportion of functionally important sites when comparing the different serotypes was significant ($\chi^2_6 = 14.034$, P = 0.029). Similarly, the proportion of structurally important sites was also significantly different among serotypes ($\chi^2_6 = 20.948$, P < 0.01). Comparison of the number of sites under functional and structural constraints between serotypes uncovers an interesting pattern that suggests certain correlation between the selective constraints acting on VP1 and the functional/structural selection shifts among serotypes. In fact, those serotypes that have not undergone adaptive evolution present the greatest number of functionally and structurally

important regions (140 and 145 sites for SAT3 and C serotypes, respectively). The percentage of sites under functional constraints as detected by CONSEQ was not an artefact of the amino acid variability produced by adaptive evolution since all positively selected regions were detected to be functionally/structurally important. Thus, high variability due to positive selection was not leading to an underestimation of functional/structural important sites.

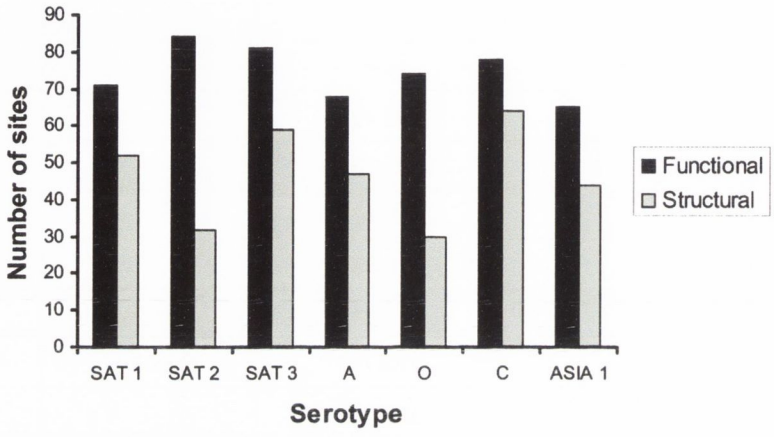


Figure 2.1: Number of functional (black bars) and structural (grey bars) important sites in each serotypes as predicted by the neural-network analysis implemented in the program CONSEQ.

2.4.3 Functional divergence between FMDV serotypes

FMDV serotypes were compared to each other and regions under functional divergence type I determined. Only amino acid sites showing posterior Bayesian probabilities greater than 0.95 were considered as having been subjected to functional divergence. In all the pairwise comparisons between serotypes, most of the sites involved in functional divergence were found in the carboxy-terminal half of the molecule, affecting antigenic sites A, C and D. Functional divergence type I between serotypes followed a trend in agreement with the geographical pattern observed in the analysis of coevolution and functional/structural sites distribution. Functional divergence affected a significantly greater number of sites when the serotypes compared belonged to different-continent locations ($t_{19} = 2.538$, $P = 0.02$; Table 2.2). Moreover, widespread viruses (O, C and A) and that in the Asian continent (Asia 1) presented lower functional divergence between each other compared to the number of sites under functional divergence when serotypes confined to the African continent (SAT1, 2 and 3) were compared ($t_7 = 3.789$, $P = 0.007$; Table 2.2).

Table 2.2: Hemi-matrix of the number of sites under functinal divergence type 1 among FMDV serotypes

Serotype	SAT1	SAT2	SAT3	ASIA1	C	A	O
SAT1	-						
SAT2	11	-					
SAT3	14	11	-				
ASIA1	13	22	18	-			
C	6	21	30	7	-		
A	15	33	80	7	5	-	
O	15	14	22	6	0	7	-

The distribution of sites under functional divergence also changed depending on the serotypes compared. Antigenic sites A, C and D were massively affected by functional divergence when African serotypes were compared (Figure 2.2). Conversely, comparison of serotypes A and O showed functional divergence in regions outside the antigenic sites, with only antigenic site A showing a moderate level of functional divergence (Figure 2.2). Comparison of serotypes belonging to completely different geographical areas showed functional divergence in antigenic and non-antigenic regions, probably indicating ancestral functional divergence.

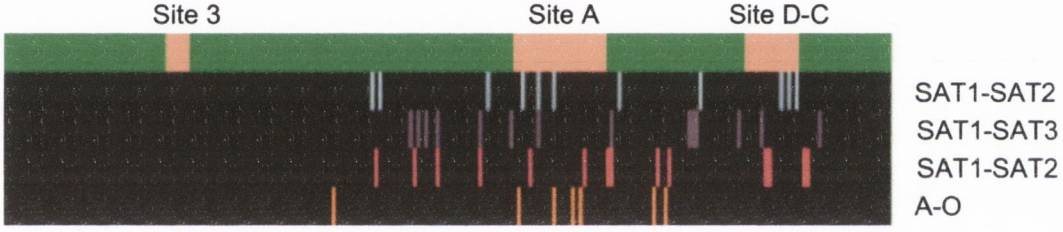


Figure 2.2: Analysis of functional divergence between FMDV serotypes. The green horizontal bar represents the sequence of the VP1 protein of FMDV, with antigenic sites pink-labelled. Vertical bars in each horizontal lane indicate the approximate location of sites under functional divergence in the serotypes compared. The different comparsions are colour-coded for clairty.

2.4.4 Intra-molecular co-variation generates FMDV immune-escape mutants

To determine if specific residues are important for the antigenicity of the virus through their functional/physical interaction with antigenic sites, we conducted an amino acid covariation analysis developed by Korber et al. (1993). This method is implemented in the program PIMIC (Codoner, Elena and Fares, unpublished and available on request). Sites detected as coevolving under this approach were, in all the serotypes, located in the antigenic sites A, C, D, and antigenic site 3. There was also a strong correlation between those sites co-evolving and those under positive selection. In fact, the percentage of co-evolving sites under positive selection were 100% (i.e. all sites under co-evolution were positively selected) in SAT2, 83% in A, 63% in Asia1, 33% in SAT1 and 21% in O (Figure 2.3). Results were not biased by the fact that

positively selected sites are more prone to be detected as coevolving because in some serotypes sites under positive selection were not detected as coevolving and vice versa. Interestingly, a pattern emerges that correlates the geographic distribution of the serotype with the molecular distribution of sites under adaptive evolution or co-evolution. For example, for those serotypes widely distributed (O and A) or covering an overlapping geographical area (O, A and Asia1) several sites under adaptive evolution are found within the same group of coevolution (Figure 2.4A). In addition, several of the amino acid sites detected by SWAPSC to be under accelerated rates of evolution were detected as co-evolving with sites under positive selection. When adjacent sites to these accelerated amino acid residues were analysed by CONSEQ, all of them were found to be functionally important. For example, the accelerated amino acid site 3 in serotype O is coevolving with the positively selected site 4, both belonging to the coevolution group G5 (Figure 2.4A). This site is in physical contact (less than 4Å distant) with functionally important regions as predicted by CONSEQ (Figure 2.4B). The same holds for amino acid site 143 that is coevolving in several groups of co-evolution with sites under positive selection (139, 140, 141 and 142; Figure 2.4A). This site is contacting or is significantly close to the RGD motif (Figure 2.4B).

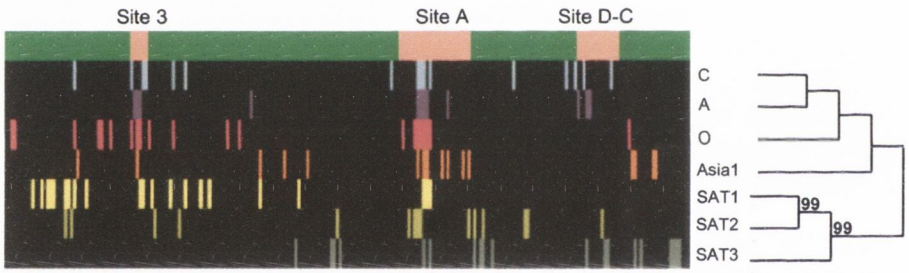


Figure 2.3: Analysis of the distribution of co-evolving sites in the different serotypes. The green horizontal bar represents the sequence of the VP1 protein of FMDV, with antigenic sites pink-labelled. Vertical bars in each horizontal lane indicate the approximate location of sites under coevolution in the serotypes compared. The different comparisons are colour-coded for clarity. The phylogenetic relationships between serotypes is also presented and internal nodes with bootstrap values greater than 75% are indicated.

Given their location, these sites are expected to be slightly deleterious and may have been thus compensated by conditional positively selected mutations as explained above. Interestingly, these sites (accelerated amino acid sites and positively selected sites) did always coincide phylogenetically further demonstrating their correlated variation. These data support thus the compensatory effect of positively selected sites in the widely distributed virus serotypes because the conditions to consider mutations as compensatory (for example, they have to coevolve with amino acid sites that have undergone accelerated rates of fixation of amino acid substitutions and these sites have to be nearby functional/structural regions) are met. Their advantageous effect however would be considered conditional to the emergence of slightly deleterious mutations

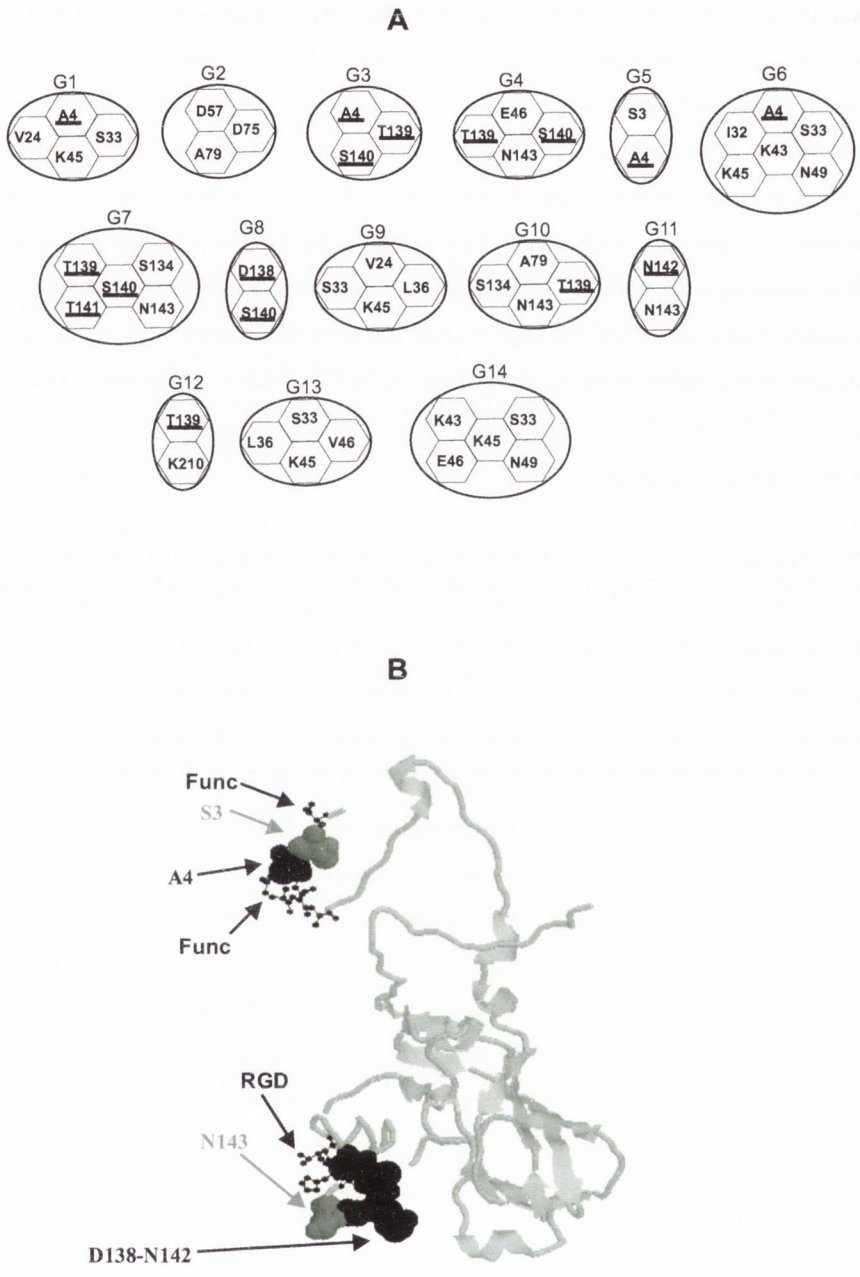


Figure 2.4: Coevolution analyses in the VP1 protein from FMDV serotype O. A) Groups of coevolution (G1 to G14) detected by the analysis. Each amino acid coevolves with all those from the same group. Underlined amino acids are those detected to have undergone adaptive evolution. B) The three-dimenstional structure of the VP1 protein of FMDV (PDB entry 1QGC). Coevolving amino acid sites are highlighted in space-fill structure, whilst functional sites are defined by the ball and stick structure. Sites in black are those positively selected; whilst sites space filled in grey represent those showing accelerated rates of evolution (where non-synonymous substitutions are significantly high but where non-synonymous-to-synoymous rates ratio is below 1). Functionally important sites (Func), predicted using the program CONSEQ, are marked. The RGD tri-peptide in the antigenic site A is also shown.

and the general effect of both types of mutations is thus neutral. There was no evidence for compensatory mutations in African serotypes. In these serotypes, sites under adaptive evolution were generally included in different coevolution groups (Figure 2.5A). No single site under accelerated rate of evolution (with no positive selection) was found to co-evolve with spatially close positively selected sites (Figure 2.5B). All four antigenic regions presented evidence of co-evolution between each other. Interestingly, the BC loop presented the highest percentage of co-evolving sites with antigenic sites A, C and D compared to other VP1 regions. A trend emerges that shows that serotype A has the greatest number of co-evolving sites in antigenic regions closely followed by C, then types O, SAT2, ASIA1, SAT1 and SAT3.

2.5 Discussion

Previous works have attempted to answer the question of whether the appearance of a new serotype is subjected to a change in evolutionary constraints. Elena and colleagues (1992) showed that a constant rate model of evolution in the different serotypes explains the evolution of three of the FMDV serotypes. They also demonstrated that episodic positive Darwinian selection might have been operating during the evolution of CS serotype. Since this statistically elegant study was performed, several new FMDV sequences were released and statistical and computational methods developed to conduct more detailed analyses of selective constraints. In this work we have conducted a detailed and comprehensive analysis of selective constraints in the different serotypes of FMDV using all the available sequences. The importance of FMDV and the amount of data available make this virus a good system to conduct a comparative evolution analysis of virus serotypes such as the one performed here. We would however like to state that other FMDV genomic regions might clarify or modify the conclusions shown in this work. In fact, previous works have shown that the FMDV host range can be governed by non-structural proteins such as the 3A and 3B proteins (Nunez et al., 2001; Pacheco et al., 2003). Analyses of other genomic regions may provide valuable information to confirm many of the conclusions in this work (Chapter 4).

2.5.1 Heterogeneous selective constraints and the emergence of immune-escape mutants in FMDV serotypes

Adaptive evolution in FMDV has been previously detected in natural isolates and has been associated with the emergence of immune-escape and persistent mutants (Tosh et al., 2003; Haydon et al., 2001; Mittal et al., 2005). Because the method used in this study examines specific branches of the tree and regions of the alignment for adaptive evolution, additional sites to those observed previously have been detected under positive selection in FMDV type

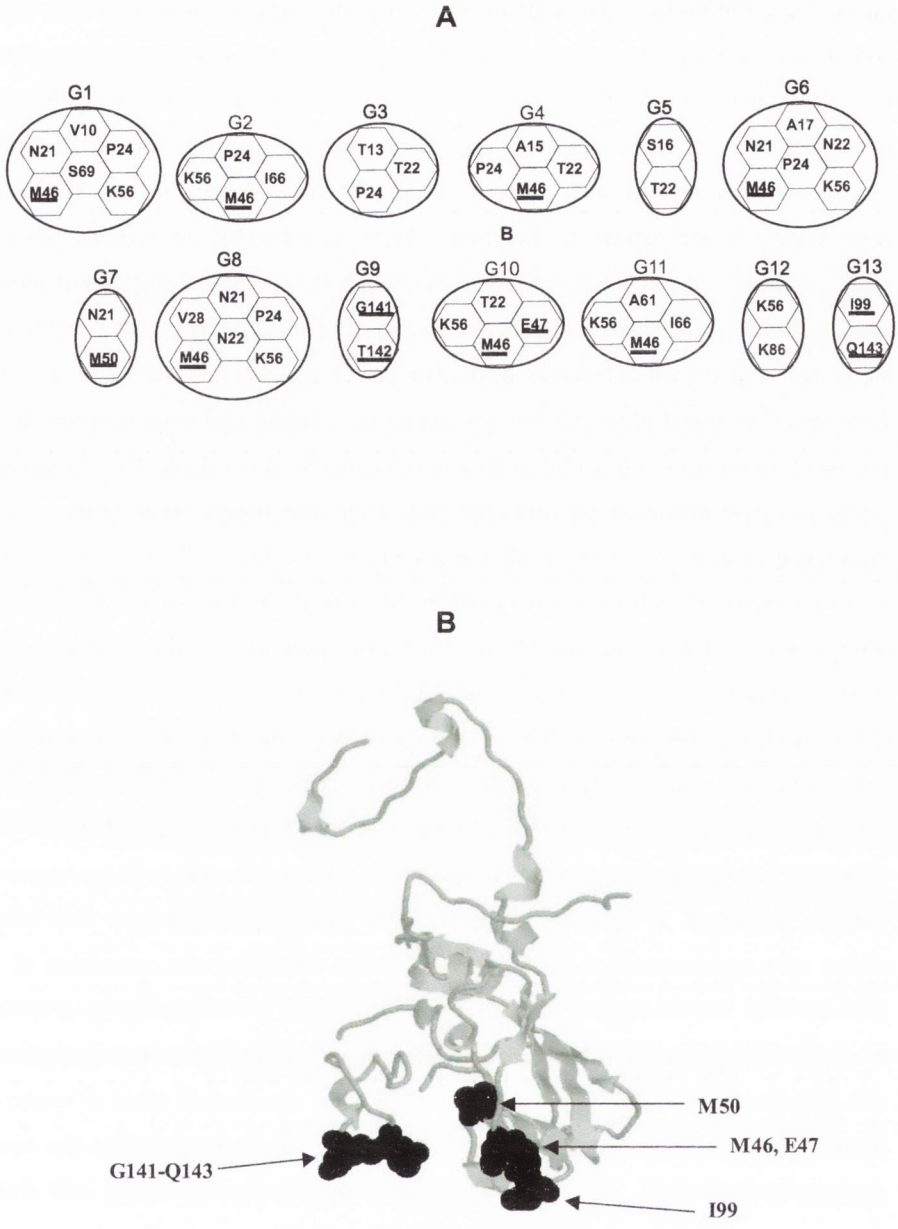


Figure 2.5: Coevolution analysis in the VP1 protein from FMDV serotype SAT1. A) Groups of coevolution (G1 to G13) detected by the analysis. All amino acids included in the same group are coevolving with each other. Underlined amino acids are those detected to have undergone adaptive evolution. B) The three-dimensional structure of the V1 protein of FMDV (PDB entry 1QGC). Coevolving positively selected amino acid sites are highlighted in space-fill structure.

A. Splitting the evolutionary time hence in as many temporal points as branches in the FMDV tree makes the non-synonymous-to-synonymous rates ratio a more sensitive approach to detect episodic changes on selective constraints. Our results are also consistent with those of Haydon et al. (2001) as we do not find any significant level of positive selection in types C and SAT3 but do detect a considerable amount in type O. Interestingly, C and SAT 3 serotypes have almost gone to extinction to this date. As a limitation of the method used here to detect selective constraints is the fact that, even though the SWAPSC software is applicable to RNA data, selective constraints operating at synonymous sites to maintain a stable secondary RNA structure may inflate ω values due to the purifying selection acting on the RNA molecule. However, this would affect all the regions in the protein and thus selective constraints would not be concentrated in few biologically important protein regions. On examination of the sites under adaptive evolution we find that G-H loop that ranges from amino acids 131-149 have undergone positive selection in all the serotypes. Within SAT serotypes we find Site 3 and antigenic region C under positive selection. Most of the regions under positive selection in SAT serotypes are not found in non-African serotypes, sparking speculation that there is a difference in the selective pressures governing the SAT types compared to non-African serotypes. On the other hand, the difference in the percentage of sites that have undergone positive selection is very noticeable, with SAT2 having a significantly greater percentage of sites undergoing selection compared to SAT1. This coincides with the fact that SAT2 is widely distributed in the African continent whereas SAT1 is confined to specific African regions, yet both serotypes overlap in some African territories. A plausible explanation for this correlation is that SAT2 has accumulated amino acid replacements in its VP1 protein that permitted the generation of immune-escape mutants and the emergence of variants in SAT1-SAT2 overlapping geographical areas as to minimize the competition by utilizing different niches in these areas (adaptive radiation). If this hypothesis were true, we would expect greater accelerated rates of evolution in the VP1 protein from serotypes in overlapping geographical areas compared to the same serotypes in non-overlapping areas. Examination of the Poisson-corrected amino acid distances between serotypes in overlapping regions reveals faster evolutionary rates for data sets of sequences in overlapping geographical regions for all the serotypes (Figure 2.6). This fact is independent of different sequence sampling dates because all the sets included sequences from the similar sampling dates.

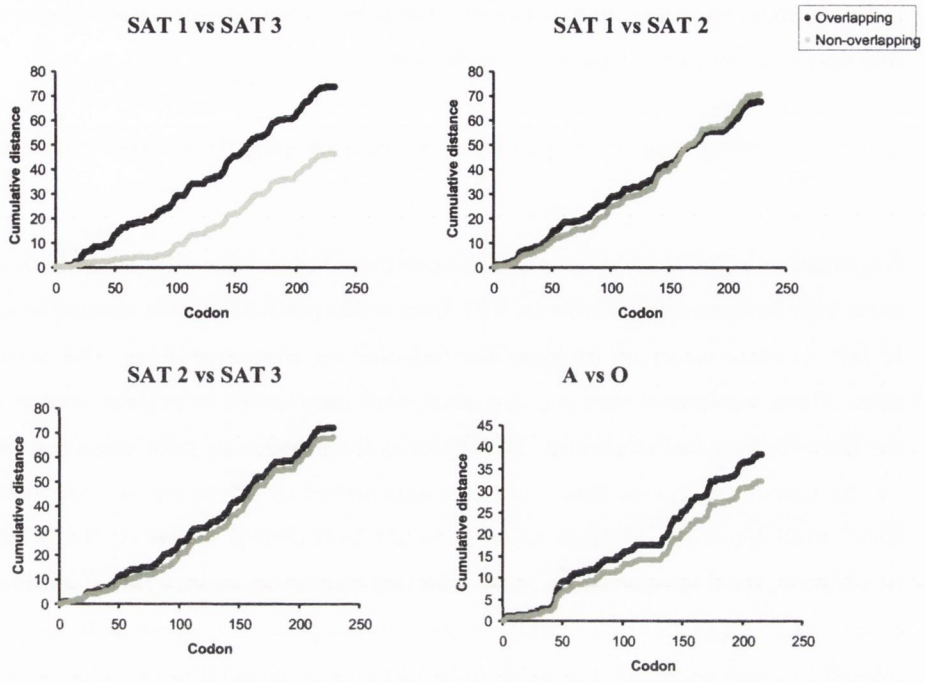


Figure 2.6: Cumulative mean Poisson-corrected pairwise distance along the sequence alignments of FMDV. Different serotypes have been compared in both overlapping geographical areas (black line) and non-overlapping areas (grey line).

2.5.2 Host-cell receptor recognition and antigenicity may be linked by positive selection and coevolution

Amino acid sites detected to be under adaptive evolution in FMDV serotypes are mostly located in antigenic regions known to interact with both, integrin receptors (Jackson et al., 1996, 2000, 2002) and neutralising antibodies (Verdaguer et al., 1995; Ochoa et al., 2000). It is remarkable the detection of positive selection in the RGD motif despite its direct involvement in the interaction with the host cell receptors. Experimental dispensability of the RGD tri-peptide for cell entry expands greatly the repertoire of antigenic variants of FMDV with substitutions at the antigenic site A (Martinez et al., 1997; Ruiz-Jarabo et al., 1999). FMDV exhibits an astonishing flexibility in receptor usage, as substitutions within the RGD or at neighbouring sites have been isolated in viruses from lesions in immune cattle challenged with a highly virulent FMDV (Taboga et al., 1997). The co-evolution of positively selected sites in antigenic regions, as shown in this work, also involved in the interaction with alternative cell surface receptors complements this conclusion. This co-evolution can be promoted by the position of the mobile loop ensuring interaction of residues in the G-H loop with residues located in distant regions of the VP1 protein such as the B-C loop (Verdaguer et al., 1999). We also concur with (Aggarwal and Barnett, 2002) who suggested that unidentified sites outside of antigenic regions may also be important

in the immune response. In this respect, our work shows non-antigenic regions under positive selection co-evolving with known antigenic sites.

2.5.3 A geographical-specific pattern of adaptive coevolution in FMDV

Examination of the three-dimensional localisation of coevolving sites suggests fixation of mutations with compensatory effects in VP1 from serotypes with a wide geographical distribution. In fact, in these serotypes we identified fast-evolving sites co-evolving with positively selected sites. These accelerated sites are in contact with functionally important sites as highlighted in the three-dimensional structures. Mutations at these sites may have important negative effects on the functional regions that must be compensated by advantageous mutations at adjacent sites. Both types of mutations together would have neutral effects on the protein's function. In addition, for those accelerated sites affecting important regions embedded in complex local structures, such as the RGD motif, several advantageous mutations at nearby sites would be needed to compensate for the deleterious effect of accelerated sites. This would explain the significant number of positively selected sites belonging to the same co-evolution group in the case of these serotypes. Conversely, those viruses with a distribution confined to the African continent showed no indication of compensatory mutations. This pattern of coevolution may have probably been produced by fixation of slightly deleterious mutations in the widely distributed serotypes (for example, serotypes O and A) subjected to population bottlenecks due to the high host geographical mobility. Caution is however stressed in interpreting these results since high host population densities may ameliorate the effects of genetic drift caused by host mobility. Conversely, SAT serotypes are confined to specific areas of the African continent probably infecting persistently the host (for example, they have their reservoir in the African buffalo) through improving their ability to generate highly competitor immune-escape mutants. In addition, compensatory mutations and the co-evolution among antigenic sites, as demonstrated in this work, would permit fixing immune-escape mutations probably maintaining the ability to recognise host-cell receptors in these serotypes. This is in agreement with previous studies suggesting the co-evolution of antigenicity and receptor usage (Mateu et al., 1996; Domingo et al., 1999; Baranowski et al., 2001). Furthermore, the fact that mutations within the B-C loop can influence the conformational stability of the G-H loop (Parry et al., 1990) supports the importance of the co-evolving sites in both loops detected here.

2.5.4 A geographical-based pattern of functional divergence between FMDV serotypes

Our analysis supports that FMDV serotypes geographically distant have diverged functionally/structurally during their independent evolution. In contrast to the African serotypes that show a geographical confinement, serotypes in the Asian continent present evidence for inter-serotype contact due to their lower functional divergence. This has been probably produced due to the greater mobility of the virus or the host that could eventually lead to bottlenecks on the viral population.

2.5.5 Virus variability and survival

In this manuscript we give details about the complex dynamics of evolution of an economically important virus. Variability in RNA viruses caused by high errors rates during virus replication seems to be important for the survival of virus populations under changing environmental conditions. It has been recently shown that low fidelity increases virus pathogenesis and that selective pressures are operating at the population level (Pfeiffer and Kirkegaard, 2005; Vignuzzi et al., 2006). The main conclusion of these studies is that pathogenesis is supported by Quasispecies diversity.

2.5.6 Sampling bias

Sampling limitations may indeed play a factor in the interpretation of these results. In order to provide accurate reflections of the disease, there is need for complete temporal, spatial coverage and full recording of all disease events. This study attempts to maximise the geographic distribution and range of sampling times for sequences in order to obtain a random and true representation of the worldwide situation rather than single episodes or epizootics. Indeed it is inevitable that the current sequences sampled here are only a subset of those actually in nature but nevertheless they aid in our understanding of the disease. In any case, this is to our knowledge the most comprehensive study done on the evolutionary constraints on FMDV serotypes.

2.6 Acknowledgements

This study was supported by Science Foundation Ireland (SFI) under the program of the President of Ireland Young Researcher Award to M.A.F. We would like to thank Prof. Santiago F. Elena for comments that have substantially improved the support for some of the conclusions and for suggesting useful additional analyses.

Chapter 3

The tale of a modern Animal Plague: Tracing the evolutionary history and timescale for Foot-and-Mouth Disease Virus

The research described in this chapter has been accepted for publication in *Virology* (Tully and Fares, 2008).

3.1 Abstract

Despite significant advances made in the understanding of its epidemiology, Foot and Mouth Disease Virus (FMDV) is among the most unexpected agricultural devastating plagues. While the disease manifests itself as seven immunologically distinct strains their origin, population dynamics, migration patterns and divergence times remain unknown. Herein we have assembled a comprehensive data set of gene sequences representing the global diversity of the disease and inferred the time-scale and evolutionary history for FMDV. Serotype-specific rates of evolution and divergence times were estimated using a Bayesian coalescent framework. We report that an ancient precursor FMDV gave rise to two major diversification events spanning a relatively short interval of time. This radiation event is estimated to have taken place towards the end of the 17th and the beginning of the 18th century giving us the present circulating Euro-Asiatic and South African viral strains. Furthermore our results hint that Europe acted as a possible hub for the disease from where it successfully dispersed elsewhere via exploration and trading

routes.

3.2 Introduction

Past epidemics of bovine spongiform encephalopathy, recurrent outbreaks of avian influenza and the latest foot and mouth disease (FMD) outbreaks in the United Kingdom have all put an increasing focus and emphasis on the origin of these veterinary pathogens. In this regard, FMD is considered to be one of the worst animal plagues endemic in many parts of the world, responsible for severe economic devastation and as such deserves special attention. FMD is a vesicular disease of domesticated and wild cloven-hoofed animals caused by a small non-enveloped virus with a positive sense RNA genome. The virus is a member of the family Picornaviridae, which contains a number of important human pathogens, such as Poliovirus, Hepatitis A virus and the common cold Human rhinovirus A. This family is considered to be among the oldest and most diverse of known viruses. Such antigenic diversity in FMDV is reflected in the existence of seven immunologically distinct circulating serotypes known as O, A, C, Asia 1 and the South African Territories (SAT) 1, SAT 2 and SAT 3.

Due to its agricultural importance FMDV has been strongly dissected from the genetic point of view yet the rate of molecular evolution of this virus or the age of the sampled genetic diversity, reflected in the time to the most recent common ancestor (MRCA), have been largely unexplored. Indeed the subject on the origin of FMDV and its respective serotypes prior to approximately 1900 remains completely unstudied. With this in mind, we have assembled a comprehensive dataset of gene sequences for all serotypes isolated at different times to infer the evolutionary time-scale and history of the disease.

The earliest commonly cited description of the disease dates as far back as 1514 to Northern Italy when an epizootic affecting only cattle was described (Fracastoro, 1546). The contagious nature of the disease was also recognised because of the observed inter-animal transmission and spread. Irrespective of this uncertainty it was not until the 17th and 18th centuries that we find trustworthy proof of its existence when FMDV affected cattle, sheep, pigs and goats from Western Europe. During the 19th century the disease had become widely diffused extending from the Caspian Sea to the Atlantic Ocean owing to the changing commercial relations between civilised countries. In 1839 the disease was first recorded in Great Britain where it quickly spread to most districts of England and some parts of Scotland (Henderson, 1978). In 1870, the virus was introduced into the United States of America from Canada as a result of diseased cattle from England. However, the disease was eventually extinguished and the U.S has remained free of FMD since 1929. It was during this time that FMDV was first recorded in South America appearing first in Argentina in 1871 after the importation of cattle by European immigrants. The origin of the disease in Africa and Asia is less well characterised but probably arose from

the transport of infected animals from Europe. Although an early description of the disease was recorded in 1780 in Southern Africa (Le Vaillant, 1795).

It was not until the early 1920s that the antigenic diversity of the disease was realised after observations by Vallee and Carre (Vallee and Carre, 1922). They showed the existence of two types designated by their areas of origin, O for the Department of Oise in France and A for Allemagne. Their work was later confirmed by Waldmann and Trautwein in Germany who named them A and B but discovered a third serotype termed C (Waldmann and Trautwein, 1926). In the 1940s, Galloway, Brooksby and Henderson, identified three additional serotypes while working at the Pirbright laboratory. These additions had been found in Southern Africa and were designated accordingly as Southern African Territories types SAT1, SAT2 and SAT3 (Brooksby, 1958). The last and seventh serotype to be discovered was Asia 1, found in the early 1950s isolated from India in 1951 and 1952 (Dhanda et al., 1957) and Pakistan in 1954 (Brooksby and Rogers, 1957).

Although descriptions of different disorders have been recorded little or no reliance can be placed on this evidence. Moreover given that this was the very first animal virus to be discovered and is undoubtedly the most economically important veterinary pathogen, a hypothesis on the origin and subsequent diversification of the virus has not been subjected to rigorous examination using gene sequence data. Similarly, the divergence times of the different types of FMD have been largely unstudied.

Thus, in order to fill these gaps in the evolutionary puzzle on FMDV and to gain a better understanding of the disease we inferred the evolutionary dynamics of the virus utilising a Bayesian MCMC framework. To achieve this, we assembled a dataset of 236 viral isolates representing all serotypes from a wide range of geographical localities and over an extensive time span. The gene analysed was VP1 because of the relatively large data sets available from previous studies (See section 2.3.1). Furthermore, the VP1 gene forms part of the structural capsid of the virion and contains several of the major immunogenic sites important to effective antibody neutralization and subsequent viral clearance by the immune system. From this, we are able to infer for the very first time the time-scale and evolutionary history of FMDV and provide direct evidence of two major diversification events, within a relatively short time span, responsible for the present day circulating strains. Indeed, our estimates yield remarkable concordance with historical accounts of the disease and highlight the importance of gene sequence data into providing new insights on the origin and spread of infectious disease.

3.3 Material & Methods

3.3.1 Viral Sequence Data

We retrieved all viral sequences with known sampling dates from GenBank (Accession numbers in addition to dates of isolation of all viruses are provided in Table 1 of supplementary information contained at `http://www.sciencedirect.com.elib.tcd.ie/science/journal/00426822`). We constructed a protein sequence alignment using the program MUSCLE (Edgar, 2004) and then we built a corresponding protein-coding nucleotide alignment based on the concatenation of nucleotide triplets. We used the VP1 gene from viruses covering a sampling period of 69 years from 1932 to 2001. In order to maximise the evolutionary information contained among serotypes we used a sub-sampling procedure, with the view that the more diverse the viruses being compared the better we can account for the different evolutionary scenarios. This was necessary due to computational constraints associated with coalescent analysis that preclude us from using large number of sequences. The original compiled dataset from which we performed our sub-sampling comprised a total of 665 FMDV sequences. The final dataset sampled consisted of 236 sequences comprising all 7 serotypes of FMDV (n): SAT 1 (32), SAT 2 (32), SAT 3 (30), A (37), O (47), C (24) and Asia 1 (34). To perform unbiased (equilibrated sampling from serotypes) sampling from the set of 665 FMDV sequences (See sections 2.3.1 and 2.3.2), we used a greedy algorithm that maximises the evolutionary divergence between sequences in a given phylogeny (Pardi and Goldman, 2005). We chose not to randomly sub-sample sequences for each serotype, as this would increase the likelihood of selecting phylogenetically close sequences thereby decreasing the phylogenetic scope contained within the serotype. In order to ensure that this sub-sampling introduced no bias into the analysis and to make robust assumptions all coalescent estimates were performed on each complete dataset independently.

3.3.2 Sequence Analysis

To determine the gene and site-specific selection pressures acting on each of the seven datasets of FMDV, we estimated the ratio of nonsynonymous (d_N) to synonymous (d_S) substitutions per site (d_N/d_S) using a maximum likelihood procedure available at the Datamonkey facility ((Pond and Frost, 2005): the single likelihood ancestor counting (SLAC) method incorporating the GTR model of nucleotide substitution, with phylogenetic trees inferred using the neighbour-joining tree method. In order to evaluate the extent of recombination we identified putative recombinant viruses in each of the datasets using the RDP3 package (`http://darwin.uvigo.es/rdp.html`) (Martin et al., 2005c) with the default thresholds. This package contains six recombination detection programs: RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), MaxChi (Smith, 1992), Chimeara (Posada and Crandall, 2001), Bootscan (Martin et al., 2005a)

and SiScan (Gibbs et al., 2000). To exclude the possibility of false positive recombination detection, we considered only putative recombinant regions detected by at least three different programs.

3.3.3 Bayesian MCMC evolutionary analyses

We estimated the rates of nucleotide substitution, the age of the most recent common ancestor (MRCA) and changing profiles in demographic histories for each geographic serotype using a model that allows for rate variation among lineages under a relaxed (uncorrelated exponential) molecular clock (Drummond et al., 2006) as implemented in the Bayesian Markov chain Monte Carlo (MCMC) method available in BEAST version 1.4.6 (Drummond and Rambaut, 2007) . We investigated four population models: constant population size, exponential population growth, logistic growth and expansion growth. In addition, we used the piecewise Bayesian skyline plot (BSP) to depict changes in genetic diversity over time. The BSP is a non-parametric method used for estimating past population dynamics through time without dependence on a pre-specified parametric model. In all cases we used sampling dates for each isolate as calibration points. We based all our estimates on the GTR+I+Γ_4 model, with the frequency of each substitution type, proportion of invariant sites (I), and the gamma distribution of among-site rate variation with four rate categories (Γ) estimated from the empirical data (parameter values available from the authors on request). In all cases statistical uncertainty in parameter values is reflected in the 95% highest probability density (HPD) values. Each MCMC analysis was run for sufficient time to ensure convergence of all parameters (ESSs >100) with a discarded burn-in of 10% (as assessed using the the Tracer program (version 1.4) (`http://beast.bio.ed.ac.uk/`). We calculated the Bayes factor (BF) to compare the performance of any two Bayesian models for the same dataset. This factor is the ratio of the marginal likelihoods with respect to the prior. This is a simple method that computes the BF via importance sampling using the harmonic mean of the sampled likelihoods (with the posterior as the importance distribution) (Suchard et al., 2001). A BF of >20, or a ln BF of >2.99, is defined as strong support for the favoured model. Although estimating marginal likelihoods is difficult so the estimates of BF outputted by Tracer v1.4 must be stated with caution. Consequently, we inherently took into account the error associated with calculating the BF using the harmonic mean approach.

3.4 Results

The maximum clade credibility trees produced from MrBayes (`http://www.sciencedirect.com.elib.tcd.ie/science/journal/00426822`) and BEAST reveal that FMDV can be divided into two groups, with distinct viral lineages observed in Africa and in Europe/Asia. Serotypes

A, C, O and Asia 1 all cluster together and are supported by high posterior clade probabilities (p>0.94). Sequences from Africa also formed a distinct monophyletic cluster that diverged from the Euro- Asiatic serotypes, and further sub-divided into three sub-clades (SAT 1, 2 and 3) (Figure 3.1). To evaluate the timescale of the divergence of serotypes we examined evolutionary rates and dates of divergence using a Bayesian coalescent approach.

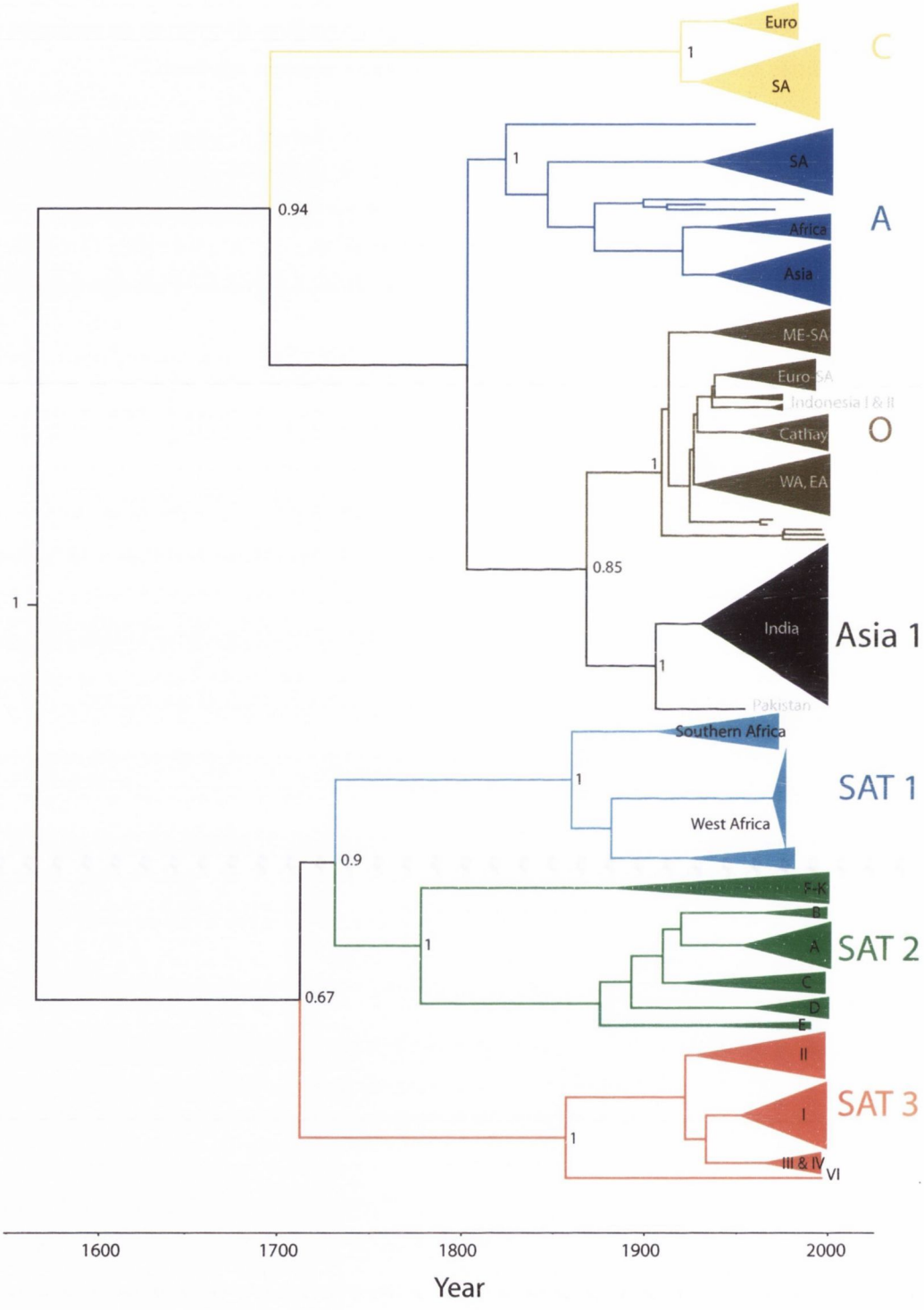


Figure 3.1 Legend on following page

Figure 3.1: Maximum Clade Credibility Tree based on 236 FMDV gene sequences. The seven serotypes are schematically represented by different colour triangles and accordingly labelled with a timeline of divergence at the bottom. In addition, we show only at respective ancestral serotype nodes those posterior probabilities greater than 0.5. Further genetic diversity within serotypes is shown, in accordance with the FMD topotype concept. For consistency names are kept in agreement with previous designations. Serotype C, contains two distinct nodes comprised of viruses, of European and South American origin, termed Euro and SA, respectively. For type A there are three distinct clades comprising isolates from South America (SA), Africa and Asia. Type O also contains a number of distinct lineages named Europe-South America (Euro-SA), Middle East-South Asia (ME-SA), Cathay (an ancient name for China and east Tartary), west and East Africa (WA and EA). For SAT 2 viruses belonging within the South African region are designated A, C, D and K) while in West Africa two regionally distinct types are found (E and F) and the remaining viruses (B, G, H, J, K) belonging to East and Central Africa. Similarly, SAT 3 contains clades corresponding to geographically distinct regions with I -IV from Southern Africa while VI corresponds to East Africa.

3.4.1 Estimation of nucleotide substitution rates

Broadly we observed equivalent rates of nucleotide substitution in all serotypes and across all 236 FMDV sequences. The mean nucleotide substitution rate was 2.48 x 10^{-3} substitutions per site, per year ranging from 1.07 x 10^{-3} (SAT 2) to 6.50 x 10^{-3} (SAT 1) substitution per site, per year (Table 3.1). To determine if these rates were affected by episodes of positive selection acting on specific codons or lineages we calculated d_N/d_S (also known as ω). Using the single likelihood ancestor counting (SLAC) method, available at the Datamonkey facility (Pond and Frost, 2005). The analysis revealed little evidence of positive selection, and an abundance of negatively selected sites. Overall, across all serotypes, we only found evidence of positive selection at a significance confidence level of $p < 0.1$ for three amino acid sites: one site for serotype A and two for serotype O. This number reduced to two sites when using the more stringent significance value of $p < 0.05$. All ω values were less than 0.2 indicating that most of the molecule has been undergoing strong purifying selection. Our results also show little if any recombination detected in the datasets (Table 3.1). Despite the occurrence of recombination in some of our alignments similar rates were observed in those serotypes where recombination was absent. Indeed, the mean substitution rate in SAT 1 was higher than any other serotype where recombination was present. In summary this supports previous studies that have demonstrated that recombination is largely constrained to non structural genes with very few phylogenetic incongruities observed in the structural or capsid proteins (Carrillo et al., 2005; Jackson et al., 2007; van Rensburg et al., 2002). Thus, we reveal that neither evolutionary process has significantly biased our estimates thereby not violating the assumptions of coalescent analysis.

Data Set (n)	Date range of sequences	Demographic Model	Substitution Rates[a]	Age (Years)[a]	d_N/d_S	Recombination detected
A (37)[b]	1932-2001	Constant	4.26×10^{-3} (2.46×10^{-3} - 6.16×10^{-3})	178 (78, 189)	0.155	Yes
O (47)[c]	1958-2000	Exponential	3.14×10^{-3} (1.84×10^{-3} - 4.34×10^{-3})	92 (72, 173)	0.124	No
C (24)	1953-1994	Logistic	1.63×10^{-3} (6.17 x 10-4 - 2.50×10^{-3})	82 (61, 145)	0.167	No
Asia 1 (34)	1954-1999	Constant	6.32×10^{-3} (3.39×10^{-3} - 9.47×10^{-3})	96 (54, 161)	0.178	Yes
SAT 1 (32)	1962-1981	Constant	6.50×10^{-3} (2.03×10^{-3} – 2.84×10^{-3})	141 (73, 255)	0.143	No
SAT 2 (32)	1948-2000	Logistic	1.07×10^{-3} (4.90×10^{-6} - 1.14×10^{-3})	223 (87, 253)	0.105	Yes
SAT 3 (30)	1965-1999	Constant	2.58×10^{-3} (1.17×10^{-4} - 5.59×10^{-3})	144 (76, 316)	0.156	No
All data (236)	1932-2001	Constant	2.48×10^{-3} (1.69×10^{-3} - 3.31×10^{-3})	432 (218, 1250)	NA	NA

Table 3.1: Bayesian estimates of divergence times and evolutionary parameters for FMDV serotypes. [a]For all parameter estimates the 95% HPD values are given in parenthesis. [b] Significant evidence for positive selection at codon 171 under the single likelihood ancestor counting method. [c] Significant evidence for positive selection at codons 48 and 139 under the single likelihood ancestor counting method. NA, not applicable.

3.4.2 Divergence times of FMDV serotypes

From the Bayesian coalescent analysis, the deepest node on the FMDV phylogeny corresponded to a time of origin with a mean age of 432 years (95% HPD of 218-1250 years, Table 1). Our analyses resulted in the description of two primary clades that diverged from a common ancestral FMDV. The first clade comprises four sub-clades encompassing isolates from serotypes A, O, C and Asia 1 entitled the Euro-Asiatic clade. The estimated mean divergence time of the Euro-Asiatic clade was at approximately 306 years ago (95% HPD of 121 and 477 years, respectively) (See figure 3.1 and Table 3.1). The oldest estimated serotype within this primary clade is that of type A. The common ancestor of serotype A isolates existed around 1823 (95% HPD intervals of 1811-1923). The second primary clade comprises those viruses indigenous to the African sub-continent known as the SAT type viruses. Similarly, with the Euro-Asiatic clade, the estimated mean divergence time of the African clade was approximately 289 years ago (95% HPD of 142-596 years).

3.4.3 Demographic history of FMDV

We found the best-fit model of population growth for each serotype under a relaxed molecular clock model allowing for rate variation implemented in BEAST v 1.4.7 (Table 3.1). We investigated four population growth models, constant, exponential, logistic and expansion growth. We calculated approximate marginal likelihoods of the four different demographic models. For serotypes A, Asia 1, SAT1 and SAT3 the Bayes Factor (BF) favoured a model assuming a constant population size. Although for serotype A the exponential population growth model had a slightly higher likelihood than the constant population size but given the standard errors associated with each model there were insufficient grounds to reject the latter model. By contrast, serotypes C and SAT 2 show a significant logistic trend, in which an initially rapid growth phase is followed by a slowdown in the growth rate towards the present. While the BF was significant (log10 BF = 5.04) for serotype O when an exponential model of population growth was assumed. The reconstruction of the demographic history of the most widely distributed serotype, type O revealed that the virus experienced a sharp increase after the 1920s until it reached its peak in the 1960s from where it started to decline gradually. In fact, this decline did not last long (10-15 years) and the trend towards the present shows a modest expansion. The Bayesian skyline plot analysis of FMDV isolates (Figure 3.2) indicates that the onset of growth started from the early 1900s and coincides with the appearance of a number of major clades in the phylogenetic tree. This graphical representation suggests that the virus grew rapidly until the 1970s, when it experienced a rapid sharp drop in population size followed by a sudden period of exponential population growth. However, comparisons of population dynamics models did not classify complex population models as significantly better fit to the data than a constant

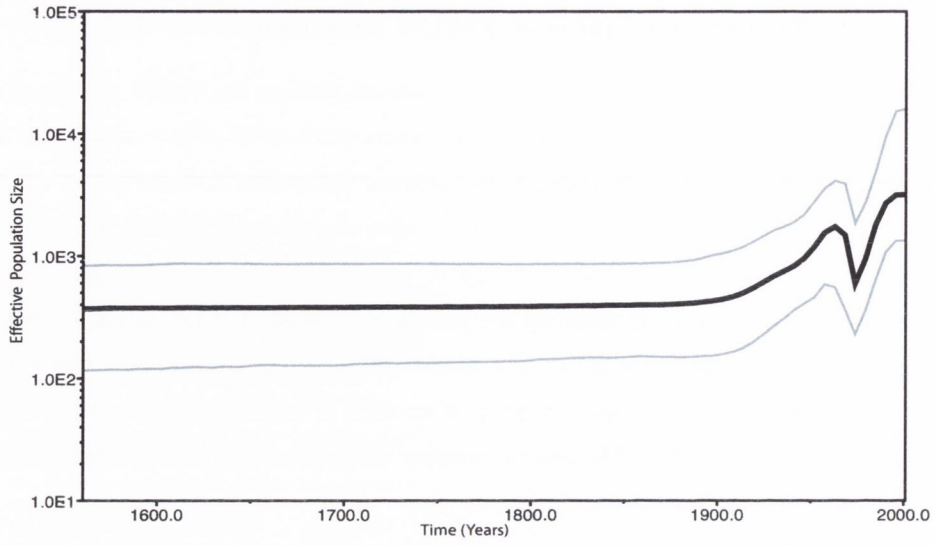


Figure 3.2: Bayesian Skyline Plot estimated from the entire dataset of FMDV isolates sampled between 1932 and 2001. The bold line represents the median estimate of the effective number of infections through time. The cyan lines indicate the upper and lower bounds of the 95% highest posterior density.

size population model (log10 BF < 3.0).

3.5 Discussion

To our knowledge, this is the first study dating the divergence times of FMDV from the Aphthovirus genus within the Picornaviridae family using modern phylogenetic techniques. The use of such techniques has proved in the past to be extremely valuable when applied to a broad range of viruses. For instance, in reconstructing the emergence of HIV/AIDS in the Americas (Gilbert et al., 2007), in estimating the origin of Yellow Fever Virus into the Americas (Bryant et al., 2007), the origin of smallpox (Li et al., 2007), investigating the spread of Hepatitis C virus (Nakano et al., 2006; Njouom et al., 2006; Pybus et al., 2003b; Verbeeck et al., 2006) and in deciphering the epidemiological dynamics of Influenza A virus (Rambaut et al., 2008). By performing a Bayesian coalescent analysis using serially sampled gene sequence data, we are able to provide important insights into the evolutionary and epidemiological characteristics of this veterinary pathogen.

Our coalescent analyses demonstrate that an ancient precursor of the virus gave rise to two early diversification events taken place within a relatively short span of time (approx 20 years). This deep bifurcation gave rise to the three South African Serotypes (SAT 1, SAT 2 and SAT 3) as well as to a lineage, which was the precursor of the present day Euro-Asiatic serotypes (A, C, O and Asia 1). This major event may reflect adaptive radiation, driven by ecological

opportunity, where fast early diversification allows access to a greater diversity of resources (hosts). Therefore different clades could reflect independent adaptive radiations into different broad niches. Remarkably, the observation of separate diversification events for FMDV has been previously suggested albeit briefly using a simplistic phylogenetic approach (Dopazo et al., 1995). Although, until now this hypothesis has never been fully subjected to rigorous examination or even attempted to be dated using modern phylogenetic techniques.

The precursor of the current circulating strains of the different Euro-Asiatic serotypes underwent further diversification with three of the serotypes occurring more recently in the early 1900s. Serotype A has a much older ancestral origin dating back to approximately 1822 (or 178 years from the youngest tip of FMDV isolated in 2001). This older date of origin is of great significance as it suggests that the virus witnessed in the Americas in the late 19th century were type A because it is well conceived that European immigrants brought the disease into the continent.

Previous hypothesis have asserted imported cattle from European immigrants in the late 1860s or early 1870s as the alleged route of introduction of the disease into South America. Our results indicate that there is a strong genetic relationship between European strains and certain South American strains clearly vindicating this hypothesis. The strong European-South American relationship between type A viruses is well studied with a designated genetic lineage known as the Euro-SA topotype (Knowles and Samuel, 2003).

An almost simultaneous diversification event occurred shortly after the initial introduction of the disease into Africa giving rise to the formation of three clades appropriately named as the South African Territories (SAT 1, SAT 2 and SAT 3) where they have been maintained in sub-Saharan Africa ever since. The timing of such an event was estimated at approximately 289 years ago placing it at the start of the 18th century. This timing is consistent with historical accounts of the disease in southern Africa with the earliest observation of a disease that bears resemblance to FMDV made in 1780 (Le Vaillant, 1795). After this, several descriptions of the disease were observed at the end of the 19th century in Rhodesia, Swaziland and South Africa. At this point our estimates show us that the remaining clades (SAT 1 and SAT 2) had already diverged. Although both of these clades shared a similar time of divergence they do not appear to undergo rapid expansion with a constant population size best fitting the data. This constant population size would be consistent with a long-term persistent infection, a theme that is well studied within the SAT serotypes owing to the carrier state of African buffalo (*Syncerus caffer*) populations where FMDV can last for a number of years. The African Rinderpest pandemic that occurred at the end of the 19th century would have severely compromised FMDV diversity and during this time no reports of FMD in Southern Africa were reported. However, we believe that the SAT strains were maintained in isolated African buffalo herds but subsequent expansion did not occur till the early 1900s. Conversely, in the case of SAT 2 a logistic model of population

growth rate best fit the data with coalescent times implying an earlier time of divergence around about 1778. Interestingly, SAT 2 is the serotype most often associated with outbreaks of foot-and-mouth disease (FMD) in livestock in southern and western Africa and is the only SAT type to have been recorded outside the African continent in the last decade with incursions made into the Middle East. Estimation of nucleotide substitution rates shows that this serotype has the lowest substitution rate of 1.07×10^{-3} with a very wide range of 95% HPD values of between 4.90×10^{-6} and 1.14×10^{-3} reflected in its low ω value. Despite this wide range of estimates similar rates of nucleotide substitutions were observed from isolates in sub-Saharan Africa. All the substitution rates found in this study are consistent with previous estimates reported for FMDV and within the range reported for 50 RNA animal viruses (Jenkins et al., 2002) and of 49 other species from another study that included more slowly or rapidly evolving viruses (Hanada et al., 2004). A very elegant older study examining the genetic diversification of serotype C over a 6-decade period attempted to date the origin of the ancestor of this serotype only to be met with considerable uncertainty (Martinez et al., 1992). But after omitting European viruses isolated after a certain date a reasonable correlation was established with an extrapolated date of origin of 1897 with 95% confidence intervals of 1876 to 1912 (Martinez et al., 1992). Our estimates indicate a similar time of origin of 1918 with 95% HPD of 1856-1940, placing our mean date within 21 years of the previous study. Not only are our time estimates similar but also the rates of nucleotide substitution coincide, placing our mean estimates at 1.63×10^{-3} in comparison to 1.43×10^{-3} found in the former study.

As we show in our study, FMDV diverged from a recent common ancestor with two primary clades evolving into separate evolutionary paths and subsequently evolved in geographically discrete animal populations. Our coalescent analyses indicate that the divergence occurred within a short time-span. But based on historical records and our dates of origin we speculate that the ancient precursor of the virus belonged to Europe. In fact, based on historical events we speculate that European exploration may have played a key role in the emergence of FMDV, not only to Africa and Asia but also in the New World (Figure 3.3). For instance, the Portuguese explorer Bartholomew Diaz was the first European to see the stormy Cape of Good Hope—the Southern-most point in Africa—in 1488. Not until 1488 was permanent settlement established here by the Europeans following the founding of a refreshment station by the Dutch East Company (VOC). The Cape was approximately mid-way between Europe and India, and as such made it an ideal stopping point where trading ships could pick up fresh food and water. Populating the Cape at this time was a large indigenous group now known as the Khosian, a semi-nomadic pastoralist race herding cattle and sheep. Despite continual trade with the Europeans the Khosian regularly raided their cattle, hindering establishment of a permanent European settlement. Later slavery became part of the Dutch colonial enterprise with shiploads of slaves arriving from West Africa to serve on farms, amongst other tasks. Taking all this into

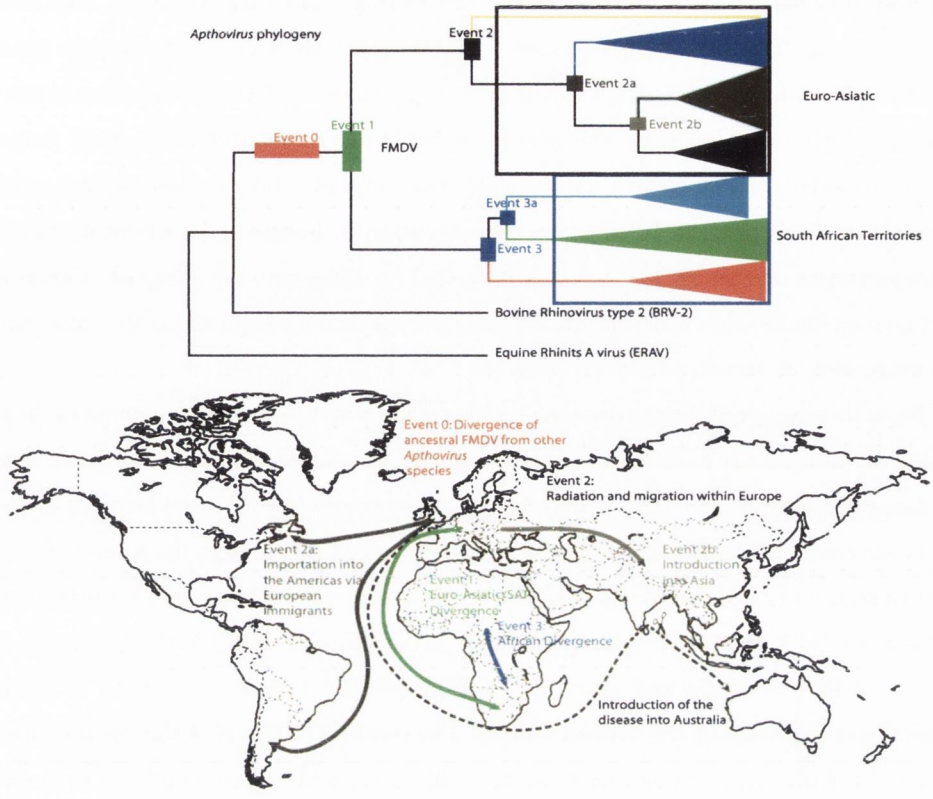


Figure 3.3: Hypothesis on the spread of FMDV from ancient times. Event 0, the clustering of FMDV with other Apthovirus species suggests an ancient diversification event from a more recent common ancestor with BRV-2 than ERAV. Note that this is purely a schematic of the topology taken from Hollister et al. (2008). Event 1, an ancestral precursor FMDV diverged into two primary clades designated the Euro-Asiatic serotypes and the South African Territories where they evolved independently. The dates and long relation of the disease in Europe suggests that the ancestral FMDV probably originated in Europe and later spread to Africa via infected animals or alternatively the divergence could have taken place in Africa. Event 2, FMDV diversification and migration throughout the Old World. Event 2a, European exploration brought FMDV into the Americas where it then spread. Event 2b, represents the diversification and migration of FMDV throughout Asia. Event 3, shows the diversification of the SATs on the African continent, followed by later introduction of additional subclades (Event 3a).

account, it would seem that there was ample opportunity for the disease to spread to various other parts of Africa, given the proximity of people working with livestock. Above we explain one of many possible scenarios for the emergence of the disease in Africa and although it may serve incorrectly in portraying the tale of this animal plague. Nevertheless it illustrates how European exploration probably played a significant role in the emergence of FMDV, not only to Africa and Asia but also in the New World. It would also explain that the early observation of the disease in Northern Italy in the 16th century were indeed accurate reports of the virus.

Given the abundance of heterochronous sequences (i.e. sequences of viral genes isolated at different times) and the apparent infrequency of recombination makes this gene a feasible and appropriate candidate to reconstruct the evolutionary pathways of FMDV. Although selection

has been previously reported within isolates (Fares et al., 2001; Haydon et al., 2001 and section 2.4.1), this carefully constructed dataset contains very little if any such selective pressure. In fact there is evidence of relatively strong purifying selection. Moreover, selection is not thought to significantly bias coalescent estimates (Grenfell et al., 2004a; Hue et al., 2005; Lemey et al., 2003). However, recombination may severely bias estimates but the similar high substitution rates observed in those serotypes where recombination is thought to be absent in comparison to those serotypes where it is reported assures us that no differences are observed. This is validated by previous studies that have shown that this protein is not subjected to the same intensity of recombination as non-structural proteins are.

Nevertheless, our ability to disentangle the reason why this pathogen emerged is hindered by ecological and genetic explanations that have been associated with anthropogenic factors such as changes in farming practices. Undoubtedly the increased level of trade between industrialised countries and other livestock movement patterns has had an effect in the spread of the disease. The emergence of the disease on the African continent was an important turning point in the history of FMDV where it is endemic with disease eradication being a grim prospect. Although the initial lack of restrictions on cattle traffic meant the disease was carried in all directions particularly by means of the African buffalo. The continual survival of this virus is likely driven by both ecological and evolutionary factors. This study adds pieces to solving the evolutionary origin of FMDV and of animal RNA viruses in general.

The implications of the results presented here will ultimately aid in our understanding of the genetic diversity between serotypes. For instance no previous study has convincingly accounted for the occurrence of lineage specific variation in rates of nucleotide substitution. The rate variation that exists between lineages within the phylogeny is suggestive of clade-specific epidemiology. In addition the reconstruction of the population history of FMDV serotypes provides indication of changing dynamics over the history of the lineage. This may give an indication of what serotypes are most likely to contribute to future epizootics. Appreciation of the underlying mutational dynamics and the ability of viruses to generate antigenic variation is undoubtedly important for epidemiological dynamics and vaccine development.

3.6 Acknowledgments

We thank Dr Valentin Ruano Rubio for his assistance with the implementation of the greedy algorithm and fellow scientists for making their software freely available. This work was supported by Science Foundation Ireland.

Chapter 4

Shifts in the Selection-Drift Balance Drives the Evolution and Epidemiology of Foot-and-Mouth Disease Virus

This chapter is based on a manuscript recently accepted in *Journal of Virology* (authors D.C. Tully and M.A. Fares).

4.1 Abstract

Foot-and-mouth disease virus (FMDV) is the causative agent of an acute vesicular disease affecting wild and domesticated animals. Despite the economic burden of the disease and all efforts to eradicate it, foot-and-mouth disease outbreaks continue to emerge unexpectedly in developed and developing countries. In-depth evolutionary genomic analyses correlating the mutational dynamics of the virus with its epidemiology remains to be performed. Using a dataset of 103 complete genome sequences representing the seven serotypes we investigate these dynamics. We first show the important role that selection plays in the genomic evolution of viral isolates for serotypes. Next we identify selection and relaxed constraints due to genetic drift through analyses of synonymous sites. Finally, we investigate interactions between mutations that show co-evolving patterns and analyse, aided by protein structural data, slightly deleterious and compensatory mutational dynamics. Specifically we demonstrate that structurally exposed capsid proteins present a greater number of adaptive mutations and relaxed selection

than non-structural proteins. Such events have been magnified during the evolution of the Southern African virus types (SATs). These shifts in selection-drift balance have generated the great antigenic and genetic diversity observed for the SAT serotypes and that are responsible for epizootics on the continent of Africa. The high number of slightly deleterious and compensatory mutations in SAT serotypes in structural proteins is also testament to such balance re-shaping. The significant accumulation of these coevolving mutations in African serotypes supports their contribution in generating adaptive immune escaping mutants and in establishing persistent infections. Interestingly, the reverse of this pattern in non-structural proteins reveals the neutral fixation of mutations in the more widely spread and commonly studied Euro-Asiatic serotypes.

4.2 Introduction

Foot-and-mouth disease virus (FMDV) is the causative agent of an acute, systemic disease that affects farm animals and therefore has economically devastating effects (Mahy, 2005). FMDV belongs to the Aphthovirus genus in the family Picornaviridae. Its positive-sense, single-stranded RNA genome is 8.5kb in length and is immediately translated into a polyprotein upon virus entry into a host cell. This polyprotein is co- and posttranslationally cleaved to yield 12 mature proteins (Rueckert, 1996; Pereira, 1991). The 12 proteins comprise, from 5´ to 3´ a set of four structural capsid proteins (1A, 1B, 1C and 1D, also known as VP4, VP2, VP3 and VP1, respectively) and 8 non-structural proteins (L^{pro}, 2A, 2B, 2C, 3A, 3B, $3C^{pro}$and $3D^{pol}$) (Porter, 1993). Structural proteins, 1A to 1D, assemble to form the capsid of the virion and, with the exception of 1A, are the first line of contact between the host cell and the virus. The interaction between both biological systems occurs specifically between the capsid protein VP1 and the RGD-dependent integrins and heparan sulfate proteoglycan receptors (Jackson et al., 2002, 2000). Non-structural proteins are involved in replicatory and other functions (Inoue et al., 2006; Moffat et al., 2005; Grubman and Baxt, 2004). 2B protein is known to be involved in membrane rearrangements required for viral RNA replication and capside assembly (Grubman and Baxt, 2004). Both non-structural proteins 3A and 3B appear to play a role in virulence and host range (Pacheco et al., 2003) with a single amino acid substitution mediating adaptation of the virus to a guinea pig (Nunez et al., 2001). The 3C protease is responsible for most of the proteolytic cleavage in the viral polyprotein and RNA replication (Vakharia et al., 1987) while the 3D protein is the core subunit of the picornavirus RNA-dependent RNA polymerase (Flanegan and Baltimore, 1977). Functions for the remaining non-structural proteins are less well understood.

The high mutation rates (Batschelet et al., 1976; Drake and Holland, 1999) and population bottlenecks experienced during the transmission of RNA viruses such as FMDV (reviewed in Escarmis et al. (2006)) account for their high genetic and antigenic variability. This genetic

variability is translated into the existence of seven immunologically diverse serotypes distributed around the world (South African Territories (SAT) 1-3, Asia 1, A, O and C) with varying degrees of selective genetic variability (Haydon et al., 2001; Knowles and Samuel, 2003 and section 2.4.1). Uncovering the selective patterns of this fixed variability is instrumental to the understanding of the epidemiological behaviour of these viruses and in monitoring possible outbreak sources. To date, most of the evolutionary studies performed in FMDV have focused on the structural surface-exposed capsid protein VP1 due to its essential role in cell-receptor recognition and escape from the immune response (Sobrino et al., 2001). Some of these studies have been aimed at describing the evolutionary parameters responsible for the emergence of specific serotypes (Elena et al., 1992). Others have yielded information on the evolutionary constraints acting on FMDV strains subjected to different experimental passage regimes (Fares et al., 2001) or naturally isolated (Haydon et al., 2001 and Chapter 2). Despite the clear role of adaptive evolution in generating variability in the VP1 capsid protein, many other factors may have shaped its evolution, including the changes in the balance between selection and drift (Chapter 2).

Most of the analyses performed in FMDV have been limited in different ways. Indeed, data have been biased by the phylogenetic nature of the analyses conducted (reviewed in Knowles and Samuel (2003)). The number of studies utilising full-length FMDV genomes for comparative evolutionary analyses has been even less significant and mainly concentrated on the intra- or intertypic comparisons of a limited number of isolates (Pereda et al., 2002; Mason et al., 2003). These studies have been also biased by the fact that only few serotypes have been considered (notably O, A and C) due to the lack of resolution of full-length genomes for some of the serotypes, especially for the sub-Saharan serotypes SAT 1 and SAT 3. The analysis of 103 complete genomes isolated from the seven different serotypes (Carrillo et al., 2005) is one of the most complete evolutionary analyses so far conducted in the comparative genomics of FMDV. However, the main objective of this study was to identify specific genomic regions under strong purifying selection. In addition, they reported novel viral genomic motifs with possible biological importance. Further, analyses of selective constraints operating in FMDV have ignored selection on synonymous sites, which is expected to be highly significant in RNA viruses due to constraints imposed over the secondary RNA structure (Chamary et al., 2006). These constraints may have led to high non-synonymous-to-synonymous rates ratios due to low synonymous substitution rates at these synonymous sites. This ratio is generally used as indicator of selection based on the assumption that synonymous sites evolve neutrally, which is generally not the case of constrained RNA molecules.

In spite of the effort invested in identifying evolutionary forces in FMDV, many questions remain to be answered. For instance understanding why serotypes differ in their evolutionary and epidemiological dynamics remains a critical research question. Herein, we are able to examine,

for the first time, how different genomic regions are under distinct selective constraints, as well as elucidating the role that shifts in the selection-drift balance plays in the emergence of variability within FMD viruses. We describe how the effects of genetic drift are compensated for in the different serotypes and genomic regions and finally reveal how the neutral fixation of mutation may provide information about epidemiological patterns of the virus. Our main conclusions are: i) Different viral types present evidence of different evolutionary dynamics throughout their phylogeny and genome; ii) These differences are mostly due to a changing selection-drift balance; iii) This changing balance has led to the differential accumulation of slightly deleterious mutations and compensatory changes between genomes and serotypes and iv) lastly we postulate that these mutational dynamics may explain the diverse epidemiological patterns observed for the different FMDV serotypes.

4.3 Material and Methods

We conducted extensive comparisons of the genomes of the seven serotypes to understand the FMDV mutational dynamics and correlate these with the epidemiological characteristics of the virus. We first identified fixation of amino acid substitutions in each one of the proteins and serotypes. Then we tested for selection and relaxed selective constraints in these proteins and serotypes by studying the accumulation of mutations at synonymous sites. We supported these analyses by the identification of slightly deleterious mutations (SDMs). Finally, we tested the shift in the balance between selection and drift by studying the phylogenetic distribution of these SDMs and by identifying compensatory mutations (CMs) through the analysis of the distribution of mutations in the structures available for FMDV proteins.

4.3.1 Genome Sequences and Phylogenetic Analyses

All complete genome sequences of FMDV were downloaded from GenBank. A total of 47, 6, 7, 27, 9, 3, and 4 full-length FMDV sequences were available for all seven serotypes A, Asia 1, C, O SAT 1, SAT 2 SAT3, respectively, collected worldwide during the period from 1951 to 2002. GenBank accession numbers for all sequences used in the study are listed from supplementary information containted at `http://jvi.asm.org/`.

We first constructed multiple protein sequence alignments for all genomes using MUSCLE (Edgar, 2004) and then we built protein-coding nucleotide sequence alignments based on their corresponding protein sequence alignments. The total length of the alignment was 7239 nucleotides. Maximum likelihood (ML) phylogenetic trees were inferred for each of the seven serotypes using the PAUP* package (Swofford, 1998). In each case, the best-fit model of nucleotide substitution was identified by MODELTEST (Posada and Crandall, 1998) as the general

reversible GTR+I+Γ_4 model, with the frequency of each substitution type, proportion of invariant sites (I), and the gamma distribution of among-site rate variation with four rate categories (Γ_4) estimated from the empirical data (parameter values available from the authors on request).

4.3.2 Identifying Adaptive Evolution in FMDV Genomes

There are two types of nucleotide mutations based on their effect on the protein amino acid composition, including synonymous (d_S) and non-synonymous (d_N) nucleotide replacements. In general, d_S accumulates neutrally because it has no effect on the amino acid composition of proteins and is hence unseen by selection, with the accumulated number of such mutations being hence proportional to time. In contrast, d_N involves an amino acid replacement and is therefore subjected to selection. The intensity of selection can then be measured by the ratio between these two rates ($\omega = d_N/d_S$). Values of $\omega < 1$, $\omega = 1$, $\omega > 1$, indicate negative selection, neutral evolution and positive selection, respectively (Crandall et al., 1999; Akashi, 1999; Sharp, 1997). However, it has been shown that stability of RNA molecule secondary structure as well as translational selection may impose constraints on synonymous sites leading to lower d_S values and consequently to inflated ω estimates (Resch et al., 2007; Parmley et al., 2006; Mayrose et al., 2007; Chamary et al., 2006). To identify selective constraints at single lineages and codon regions we used the sliding window approach (Fares et al., 2002a) implemented in the program SWAPSC version 1 (Fares, 2004). This program isolates all those genomic regions that show significantly lower d_S values than expected under neutrality and yields then unbiased estimates of ω (Fares et al., 2002a). A coding region will therefore be detected to have undergone adaptive evolution if it presents significantly greater number of non-synonymous substitutions per non-synonymous site while d_S accumulates following the neutral evolution model. Window sizes for the analyses were statistically optimized as to provide the lowest possible false discovery rate as detailed in (Fares et al., 2002a). We tested for the presence of adaptive evolution at each one of the lineages leading to each one of the seven FMDV serotypes and at each one of the FMDV proteins. We finally conducted all the statistical analyses to compare selection results between serotypes and genomic regions by normalising the proportion of codon sites under adaptive evolution.

The same dataset used here has previously been subjected to recombination analyses with the authors of that study coming to the conclusion that the capsid proteins undergo recombination infrequently while non-structural proteins may indeed undergo complex recombination events in some serotypes (Carrillo et al., 2005). Indeed, other studies have confirmed the observation that recombination has occurred more often among non-structural than structural genes (Jackson et al., 2007; van Rensburg et al., 2002). Hence we did not test for recombination given that any observed patterns have been previously described using standard methodologies.

4.3.3 Identifying Constraints at Synonymous Sites

To determine the degree of relaxed selective constraints in each one of the proteins for each FMDV serotype, we estimated d_S by the method of Li (1993) as implemented in SWAPSC. Due to the limited evolutionary signal contained in one codon site, we slid a window of 20 codons along the genome and estimated d_S for each sliding step. We then tested whether d_S was homogeneously distributed along the genome, as it should be expected under a neutral evolution model for synonymous sites. In particular, we were interested in knowing if d_S accumulate differently in structural proteins than in non-structural proteins and whether this effect is serotype-specific. Such difference could indicate differences in the rates of changes in the drift and selection forces between structural and non-structural proteins as well as between serotypes. If these differences were correlated with epidemiological or infectivity differences, then we could speculate about the role that drift has in driving the evolution and epidemiology of particular serotypes. In our analyses, we compared different genomic regions and hence differences in the time of isolation of sequence had no effect on the estimates of d_S. We performed d_S comparison of structural versus non-structural proteins because of their clearly differentiated roles. We also tested whether these differences were serotype-specific.

4.3.4 Identifying and Quantifying Slightly Deleterious Mutations

To quantify the number of SDMs we developed a new and simple method. Briefly, we first identified sites under strong purifying selection (highly conserved sites in the alignment). We then calculated the amino acid transition scores for the pair wise sequence comparisons at these sites. We calculated transition scores following the appropriate BLock Substitution Matrices (BLOSUM) (Henikoff and Henikoff, 1992) given the average pairwise distance for the alignment. Once the distribution of BLOSUM values was determined for that site in the alignment (for example the alignment containing the 103 sequences), we tested each serotype for the accumulation of SDMs by identifying lineages accumulating changes with transition values showing significant values (highly negative transition scores) when compared to the distribution of values of the rest of the alignment. The rationale behind this analysis is that highly conserved amino acid sites (for example, those amino acid sites showing BLOSUM values highly positive or close to 0) are expected to be functionally important because of their high level of conservation. A strong change at these sites in particular lineages (for example at terminal lineages within a serotype) is more likely to be a SDM fixed by genetic drift than an adaptive change. Following this procedure, we identified SDMs in each one of the proteins at lineages leading to each one of the serotypes and within serotypes. As in all the cases, we normalised the proportion of sites showing evidence of being SDMs as to make possible the direct comparison of these proportions between genomic regions and serotypes. This normalisation consisted on the transformation of

these per-gene and per-serotype proportions in the proportion taking into account serotypes and genes together.

4.3.5 Identifying Compensatory Mutations in the FMDV Proteome

We searched for conditional advantageous mutations by performing a study of the distribution of SDMs in the available crystal structures for FMDV proteins. Seven proteins have so far been crystallised with corresponding PDB identifications in brackets: 1A to 1D (Virus capsid 1QGC), 3C, (2BHG) 3D (1UO9) and the Leader protein (1QOL). For each one of the SDMs identified following the procedure detailed above, we searched for mutations showing the same phylogenetic distribution pattern. For example, if the SDM shows a phylogenetic pattern of mutation in the seven serotypes of "0000100", with 1 indicating strong amino acid change and 0 conservative change, we searched for all those mutations showing that exact pattern "0000100". Then both mutations were plotted in the crystal structure of the corresponding protein and the Euclidean distance between them calculated. We calculated this distance as the average distance between the atoms of the amino acid sites to which both mutations belong. Two mutations were considered to have compensated each other if, in addition to presenting the same phylogenetic pattern of transition scores, they were located at a distance less than 4Å from each other in the protein crystal structure. The significance of this distance is to ensure that two consecutive amino acids have their side chains contacting physically. In general the range of 4 to 8Å ensures taking into account contacting amino acids. However the shortage of the distance between two amino acids can be of 1.5Å or even less. Therefore the use of 4Å is orientative, conservative and a reasonably safe choice of interaction criteria. Furthermore it has been shown that the 4Å distance as a contact threshold has the biggest effect on the number of domain–domain interactions observed (Bolser et al., 2003). Also, two amino acid sites can compensate each other indirectly. For example, if site "A" and site "B" are at more than 8Å distance but are surrounding (at less than 4Å) an important functional site "C", then changes at site "A" may affect site "C" which has to be compensated by changes at site "B". We also considered these instances where site "A" and "B" were compensating each other. To account for this last situation, we identified those SDMs showing the same phylogenetic patterns but located at a distance greater than 4Å from each other in the protein structure. Then we identified sites between them in the structure showing very low divergence levels in comparison with the rest of the molecule. We measured divergence levels per site by estimating the Poisson amino acid distances for each amino acid site in the multiple sequence alignments. We normalised all the results as explained above to make direct comparisons between serotypes and genomic regions.

4.4 Results

4.4.1 Different Evolutionary Forces among Genomic Regions of FMDV Serotypes

Selection analysis identified differences in the degrees of selection between proteins as well as serotypes (Figure 4.1A). We detected adaptive evolution (diversifying selection) not only in the most exposed structural protein 1D but also in all of the structural proteins analyzed (Figure 4.1A). Among these proteins 1D presented the highest percentage (15%) of codon sites undergoing adaptive evolution affecting 4 out of the 7 serotypes. Proteins 1B and 1C have undergone adaptive evolution in almost all the FMDV serotypes, whereas 1A presented evidence of diversifying selection only in the African serotype SAT 3 (Figure 4.1A). Interestingly, the proportion of codon sites under adaptive evolution in structural proteins correlated with the percentage of proteins' exposed surface ($\rho_{Spearman} = 0.966$; $P = 0.025$).

In non-structural proteins, we also detected adaptive evolution affecting many serotypes but the amount of serotypes undergoing adaptive evolution was significantly reduced compared to structural exposed proteins. For example, except for the case of protein 1A, the average number of serotypes having undergone adaptive evolution was 5 out of 7 in the case of structural proteins; whereas in non-structural proteins the average was 2 serotypes. Notably, the detection of positive selection was not dependent upon the variability of the proteins because non-structural proteins 3A and 3B exhibit similar divergence levels as those shown by structural capsid proteins (Carrillo et al., 2005) and yet the percentage of sites under diversifying selection is lower than the rest of non-structural proteins (Figure 4.1A). In general, the percentage of protein sites under adaptive evolution in structural proteins ranged between 7.05% and 15.32%, whereas non-structural proteins showed percentages ranging between 3.51% and 11.11%. When accounting for serotypes, comparison of the percentage of codons under adaptive evolution showed greater mean for structural than for non-structural proteins, although the difference was not significant ($\chi^2_1 = 0.902$, $P = 0.342$). When taking into account serotypes individually, structural proteins presented higher percentage of sites under adaptive evolution compared to non-structural proteins (Figure 4.1B). The only two exceptions were serotypes A, with slightly greater percentage of adaptively selected sites in non-structural proteins, and Asia 1, which was the most represented serotypes under adaptive evolution when analyzing individual proteins (Figure 4.1A).

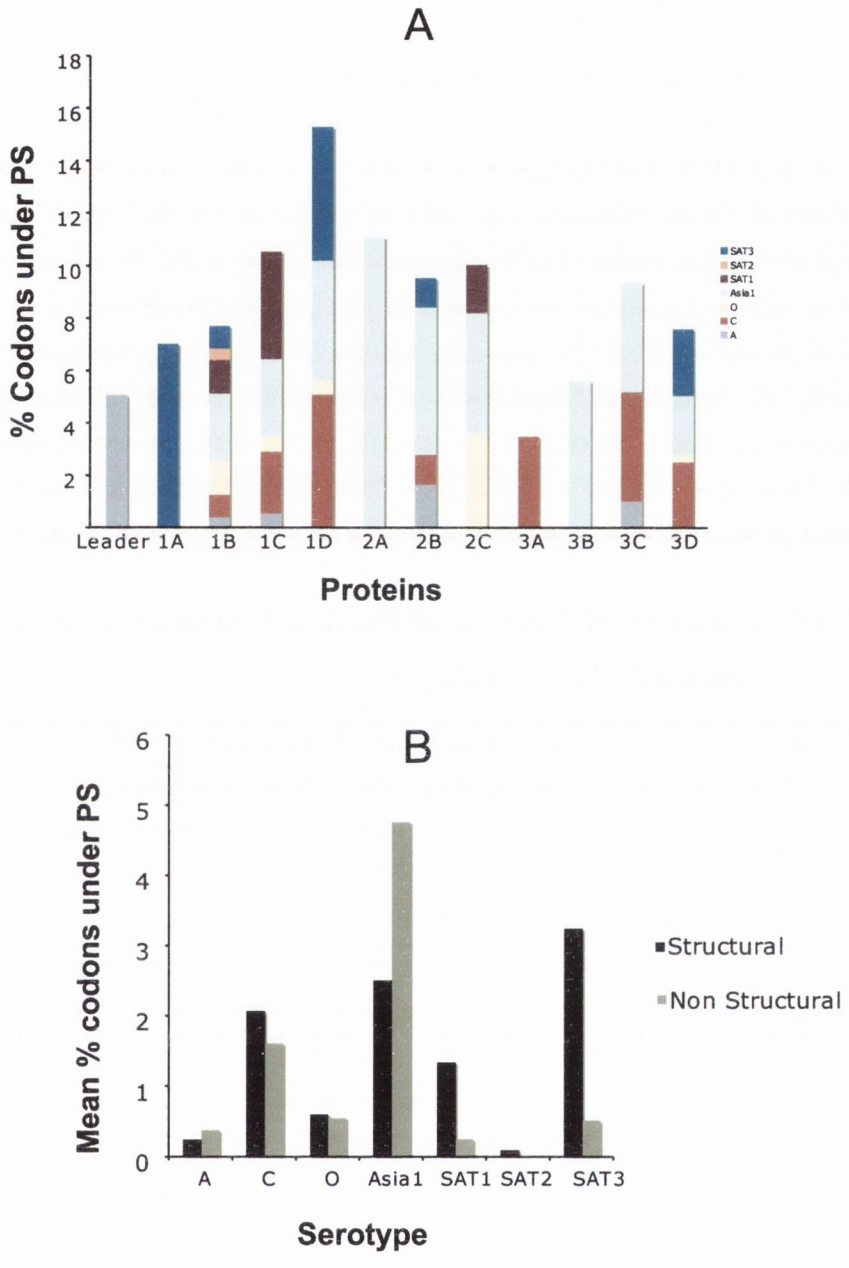


Figure 4.1: Distribution of the percentage of codon regions under adaptive evolution along the genome of FMDV for serotypes. A) Plot of the percentage of codons under adaptive evolution in the different proteins of FMDV. Different serotypes are colour-coded. B) Comparison of the mean percentage of codon regions under adaptive evolution in each serotype between structural proteins (black bars) and non-structural proteins (grey bars).

4.4.2 Relaxed Selective Pressures at Structural Capsid Proteins in SAT Serotypes

Analysis of d_S shows that its distribution was highly heterogeneous throughout the genomes. Structural capsid proteins presented significantly greater d_S values than non-structural proteins (One-way ANOVA test; $F_{[1,2331]} = 590.370$, $P << 0.001$, Figure 4.2A). In addition, we also noticed significant differences in d_S values between serotypes, with SAT serotypes presenting the highest d_S values compared to the other serotypes (Figure 4.2B). An analysis of the contribution of serotype and structural versus non-structural genes to the difference in d_S values shows that both factors as well as the interaction between them contribute significantly to the difference in d_S . For example, SAT serotypes show greater difference in d_S values compared to non-SAT types (Euro-Asiatic serotypes) ($F = 251.713$, $P << 0.001$), structural proteins show greater d_S values than non-structural ones ($F = 592.201$, $P << 0.001$) and the interaction between serotype and type of protein is also significant ($F = 92.394$, $P << 0.001$).

4.4.3 Differential Fixation of Slightly Deleterious Mutations among FMDV Serotypes

The percentage of SDMs varied along the genomes and phylogeny with most of these mutations having been fixed at the terminal phylogenetic branches of each serotype. This supports previous tests suggesting that there is correlation between the age of mutations and their adaptive value (Pybus et al., 2007). The average percentage of mutations showing evidence of being SDMs over the genomes and serotypes was 4.8% (Table 4.1). All the serotypes presented evidence of SDMs in most of the proteins except in the structural protein 1A (Table 4.1). In general, among variable amino acid sites the proportion of SDMs was higher in structurally exposed capsid proteins than in non-structural proteins, although the difference was only marginally significant ($\chi^2_1 = 3.60$, $P = 0.057$). SAT serotypes also presented significantly greater percentages of SDMs than other serotypes ($\chi^2_1 = 6.270$, $P = 0.012$). In addition, SAT serotypes presented higher proportion of SDMs at structural proteins than at non-structural proteins ($\chi^2_1 = 5.751$, $P = 0.016$). Interestingly, 1A presented no evidence for accumulation of SDMs. The next question we asked was whether SDMs are heterogeneously distributed throughout the FMDV phylogeny.

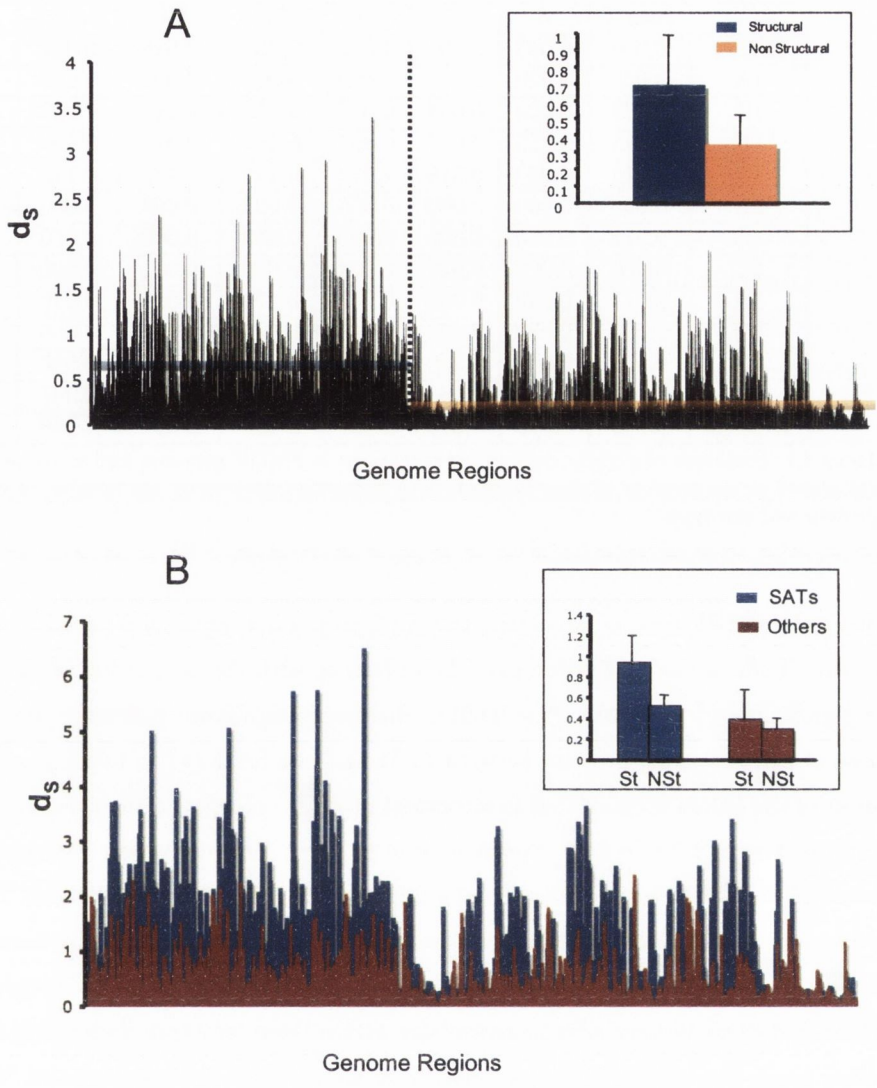


Figure 4.2: Genomic distribution of nucleotide substitutions per synonymous sites (d_S). A) average d_S variation along the FMDV genome. This distribution is highly heterogeneous when we compare structural proteins with non-structural proteins (separated in the plot by a dashed line). The median of dS for structural proteins (indicated by a blue shaded horizontal bar) is qualitatively higher than that for non-structural proteins (indicated by an orange shaded horizontal bar). The difference between the mean d_S values between structural and non-structural proteins is significant at the 5% significance level (box in the upper right hand of the plot). B) Comparison of the distribution of d_S between southern African serotypes (blue bars) and non-SAT serotypes (red bars) along the genome. These values are qualitatively greater in the SATs compared to the other serotypes along the genome. This difference is mainly due to more relaxed constraints on synonymous sites of structural (St) proteins compared to non-structural proteins (NSt) in the SAT serotypes (box in the upper right hand of the plot).

In Figure 4.3 we represent the phylogenetic distribution of SDMs in FMDV using the 103 full-length genomes. This figure highlights differences in the patterns of distribution of SDMs

Protein	Serotype							
	Asia 1	A	C	O	SAT 1	SAT 2	SAT 3	Average
Leader	0.000	0.035	0.000	0.017	0.252	0.026	0.017	0.050
1A	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1B	0.068	0.000	0.068	0.017	0.230	0.111	0.137	0.090
1C	0.015	0.015	0.029	0.037	0.292	0.212	0.109	0.101
1D	0.018	0.000	0.018	0.012	0.136	0.142	0.148	0.068
2A	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2B	0.012	0.000	0.000	0.000	0.070	0.000	0.000	0.012
2C	0.000	0.009	0.022	0.034	0.070	0.031	0.006	0.024
3A	0.011	0.040	0.009	0.017	0.070	0.029	0.070	0.035
3B	0.000	0.042	0.028	0.015	0.028	0.011	0.056	0.026
3C	0.000	0.000	0.000	0.026	0.221	0.065	0.000	0.045
3D	0.109	0.156	0.063	0.086	0.281	0.102	0.047	0.121
Average	0.019	0.025	0.019	0.022	0.138	0.061	0.050	0.048

Table 4.1: Fractions of slightly deleterious mutations in FMDV genomes and serotypes. Fractions were calculated as the number of slightly deleterious mutations divided by the number of mutations in each protein and serotype.

along the FMDV genome and along the phylogeny. First, structural proteins accumulate most of the SDMs in the SAT serotypes (the difference with the proportion of SDMs in non-SATs is significant; $\chi^2_1 = 10.702$, $P = 0.001$), whereas no significant difference was observed in the case of non-structural proteins between SATs and non-SATs ($\chi^2_1 = 1.095$, $P = 0.295$). Second, most of the SDMs accumulated in structural proteins occurred in the ancestor of the serotypes (Figure 4.4a and Table 4.2). Non-structural proteins showed however the opposite trend with most of the SDMs accumulating in the terminal branches (Figure 4.4b, and Table 4.2). This pattern indicates that structural proteins may have accumulated slightly harmful amino acid substitutions at the basal lineages of the serotypes especially in SAT serotypes, and some of these mutations became advantageous later within these serotypes undergoing fixation. On the other hand, non-structural proteins may have accumulated non-compensated SDMs.

Serotypes	Structural Proteins		Non-Structural Proteins	
	Root	Terminal	Root	Terminal
SATs	3.68 ± 1.64	0.81 ± 0.13	0.08 ± 0.09	1.21 ± 0.85
Euro-Asiatic	0.46 ± 0.33	0.11 ± 0.07	0.00 ± 0.00	0.66 ± 0.23

Table 4.2: Distribution of the percentage of SDMs in the different phylogenetic levels in SAT types compared to Euro-Asiatic serotypes. All percentages were normalised between serotypes and genomic regions as to make them comparable.

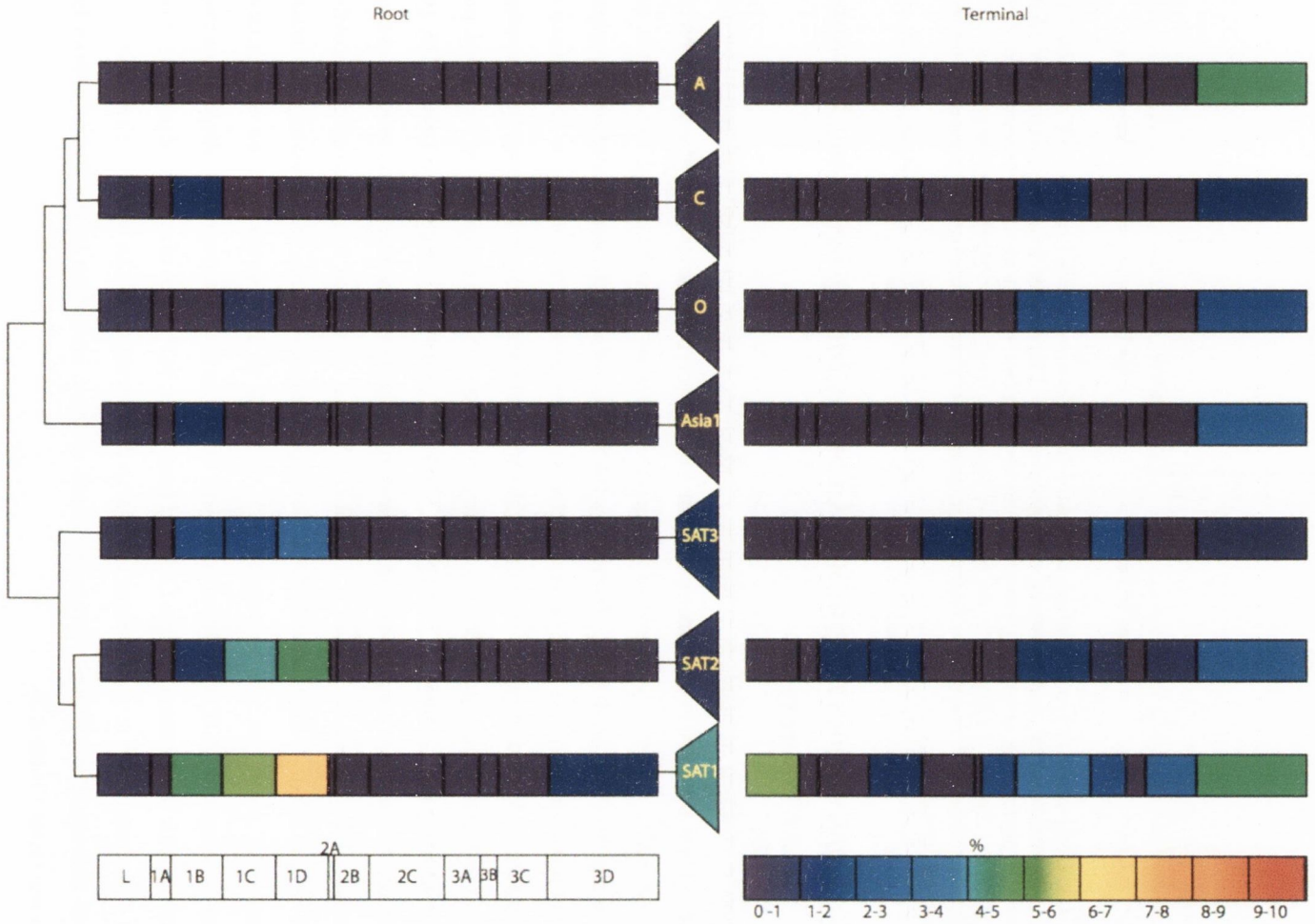


Figure 4.3: Phylogenetic distribution of the percentages of slightly deleterious mutations (SDM) in each serotype. Horizontal bars represent the genome of FMDV. Horizontal bars in the lineage leading to each serotype as well as in the tip of each serotype indicate percentage of SDMs in the ancestral serotype lineage and within serotypes, respectively. The percentages of SDMs are normalised along the genome and phylogeny and are therefore comparable between serotypes, between proteins or even between temporal sampling in the phylogeny (ancestral versus terminal distributions). The normalised percentages of SDMs are colour coded, as is the serotype group according to the average percentage of SDMs observed per serotype.

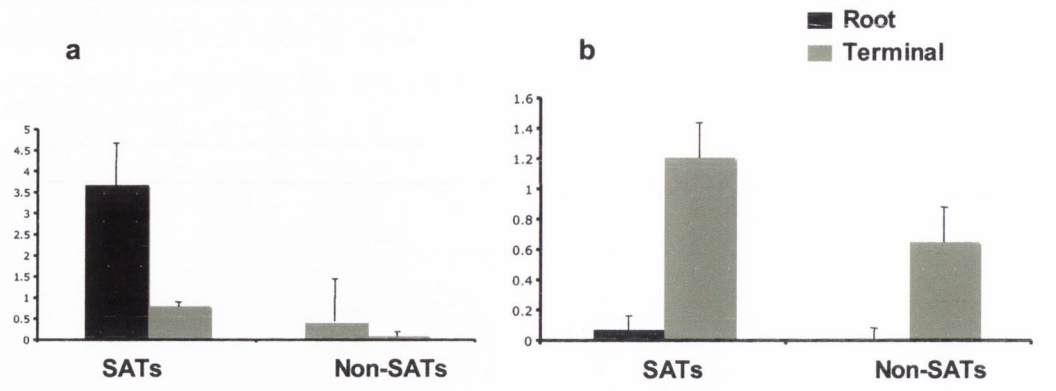


Figure 4.4: Comparison of the percentage of slightly deleterious mutations between SAT and non-SAT serotypes in structural (a) and non-structural proteins (b). Percentages have been normalised along genomes and phylogenies as to make them comparable. a) Black bars refer to the percentage of SDM identified in the lineages leading to the ancestors of serotypes (root of serotype), whereas grey bars refer to those percentages within serotypes (terminal branches). Standard errors have been averaged for root and terminal lineages.

4.4.4 Compensatory Mutations Generate Immune-escaping Mutants in SAT Serotypes

We considered a pair of SDMs to compensate each other if they presented evidence to be at a distance of less than 4Å in the protein structure or alternatively directly surrounding a highly conserved amino acid site (assumed to be conserved because of its functional importance, see Material and Methods for details). To measure the amount of advantageous mutations compared to SDMs fixed for each one of these two patterns we searched for the existence of evidence of compensatory mutation events in each one of the serotypes and lineages following the procedure described in Material and Methods. The mean percentage of SDMs compensated for in each one of the analysed proteins were 13.7%, 32.83%, 20%, 21%, 30.9% and 9.03% for Leader, 1B, 1C, 1D, 3C and 3D proteins respectively. In general, structural proteins showed on average greater percentage of compensated SDMs (24.61% of the SDMs) compared to non-structural proteins (17.9% of the SDMs). Furthermore, SAT serotypes presented more compensatory mutations than non-SAT serotypes. An average of 56% of SDMs in SAT serotypes were compensated whereas only around 7% of the SDMs in non-SAT serotypes were compensated. We also analysed the phylogenetic distribution of these CMs and found that, in the case of structural proteins, most of the CMs took place in the lineages leading to the ancestors of the south-African serotypes whereas non-structural proteins presented most of these mutations in the terminal branches within serotypes (Table 4.3).

Table 4.3: Complete list of all compensatory mutations detected in each serotype and in each gene with respect to its location in the root or in the terminal branches of the FMDV phylogeny. Each pair of compensatory mutations is enclosed within brackets while pairs are separated by a semi-colon (;). Numbers indicate the position of that site in the three-dimensional structure of the corresponding protein. When no compensatory mutations have been found we indicate it with (-).

Serotype	**Protein**	**Root**	**Terminal**
Asia 1	1B	(125, 107); (203, 201); (203, 107)	-
	3D	-	(441, 444)
A	-	-	-
C	1B	(125, 82); (125, 143); (81, 82)	-
O	1C	(186, 187)	-
	3D	-	(210, 88)
SAT 1	Leader	(127, 118); (127, 126); (127, 128); (71, 78); (71, 82); (71, 70); (118, 126); (118, 119); (118, 180); (118, 128); (118, 130); (109, 106); (78, 82); (78, 70); (67, 82); (67, 56); (67, 70); (82, 56); (82, 70); (126, 180); (119, 128); (119, 130); (180, 128); (180, 130); (97, 87); (188, 192); (128, 130)	-
	1B	(18,19); (197, 198); (197, 112); (197, 113); (197, 189); (114, 113); (198, 112); (198, 154); (168, 209); (168, 166); (168, 171); (112, 154); (112, 113); (21, 19); (41, 40); (166, 171)	-
	1C	(16, 15); (194, 197); (194, 191); (194, 193); (194, 192); (194, 111); (175, 177); (197, 193); (197, 192); (197, 111); (81, 82); (81, 80); (81, 83); (82, 80); (82, 83); (191, 192); (178, 177); (80, 83); (193, 192); (87, 83)	-

Serotype	Protein	Root	Terminal
	1D	(18, 16); (18, 17); (16, 17); (16, 14); (175, 80); (175, 173); (106, 107); (106, 82); (106, 105); (76, 117); (17, 14); (82, 170); (82, 80); (90, 167); (98, 87); (98, 100); (170, 173); (167, 87); (87, 100); (87, 105); (173, 111); (100, 105); (40, 177); (299, 303); (405, 401); (305, 306); (305, 303)	-
	3C	-	(16, 165); (139, 107); (176, 107);
		-	(108, 107); (113, 132)
SAT 2	1C	(86, 85); (165, 168); (165, 167); (168, 181); (168, 167); (104, 209); (181, 180); (121, 74); (121, 189); (121, 120); (176, 180); (189, 120);	-
	1D	(125, 74); (89, 167); (167, 87); (87, 105)	-
	3C	-	(118, 128); (327, 320); (167, 34)
	3D	-	
SAT 3	1B	(138, 136); (76, 182)	-
	1D	(86, 168); (86, 170); (168, 173); (170, 173); (97, 100)	-

4.5 Discussion

Previous studies have focused on identifying the evolutionary forces operating at particular protein-coding genes mostly involved in host-cell receptor recognition and evasion from the host immune response. Some of these studies aimed at identifying the origin of specific serotypes of FMDV such as the CS serotype (Elena et al., 1992). We have previously shown that selective constraints operating within the VP1 protein are heterogeneous among the seven FMDV serotypes and are due to the differential action of adaptive evolution (Section 2.4.1). This observation was in agreement with previous results in natural as well as in experimental isolates when using a more limited number of sequences (Haydon et al., 2001; Fares et al., 2001). In

addition, because adaptive evolution affected the antigenic regions in VP1, we speculated that genetic variability was associated with the emergence of immune-escape and persistent mutants that were previously observed (Haydon et al., 2001; Tosh et al., 2003; Mittal et al., 2005). Co-variation between the different amino acid sites within and between antigenic regions has been also linked to the ability of FMDV to escape from the immune response (Section 2.4.4). Only a handful of evolutionary analyses have examined the evolutionary constraints in regions other than the VP1 protein or its precursor, P1 (Mason et al., 2003; Carrillo et al., 2005, 2007). However, variation of adaptive evolution between genomic regions and serotypes has not been previously examined.

4.5.1 Diversifying Selection Has Driven the Evolution of FMDV Genomes

Our results indicate that most of the FMDV genomes show some evidence of the accumulation of adaptive mutations. These results are in contrast to a previous report that showed protein 1D to be the only structural protein under adaptive evolution (Carrillo et al., 2005). These differences however may be due to the fact that they failed to screen for lineage-specific adaptive evolution. Furthermore, the lack of signal for adaptive evolution in 1A in most serotypes may be due to its conserved nature and to its important role in the conformational changes of the capsid upon binding of cell receptors as well as in the generation of ions that allow RNA entry into the cell (Hogle, 2002). In addition, 1A is the only structural protein with no exposed surface and hence has no role in generating immune-escape mutants. The effect of the percentage of surface exposed amino acids in structural proteins and their importance for the biology of FMDV can be glimpsed from the positive correlation between the absolute proportion of codons under adaptive evolution and the amount of exposed surface we observe. This correlation may also reflect the evolutionary dynamic of capsid proteins in relation to its ability to interact with the host cell.

All non-structural proteins were detected to be under adaptive evolution, although the fact that only few numbers of serotypes showed evidence for selective forces may explain the previously undetected selection in these proteins when all serotypes were analysed together (Carrillo et al., 2005). Among the serotypes, type A is the one showing the lowest proportion of codon sites under adaptive evolution. FMD caused by serotype A is endemic in India, where this serotype co-exists with types O and Asia 1. This serotype is the most divergent (Tosh et al., 2003, 2002) and has been responsible for several outbreaks in India in recent years (Mittal et al., 2005). However, adaptive evolution in this serotype has occurred in the protease 3C and in the transmembrane protein 2B, but very modestly in the capsid exposed proteins (showing very low percentages of codons under adaptive evolution in only proteins 1B and 1C, but not in 1D). 3C protease is involved in RNA replication and processing of the polyprotein and is considered an important drug target (See for example (Sweeney et al., 2007; Curry et al., 2007)) Further, pro-

tein 2B is involved in membrane rearrangements required for viral RNA replication and capside assembly (Grubman and Baxt, 2004). Hence, these results lend support to the hypothesis of selection for faster replication in this serotype, which may generate spontaneous outbreaks.

4.5.2 Varying Selective Constraints at Synonymous Sites between FMDV Genomic Regions

The main conclusion resulting from the analysis of synonymous nucleotide substitutions is that these are heterogeneously distributed along the genome as well as between serotypes. The enhanced heterogeneous genomic distribution of dS values in SAT compared to Euro-Asiatic serotypes pinpoints the complex evolutionary dynamics of the different FMDV proteins and the possible dependency between these dynamics and the epidemiological behaviour of the different serotypes. SAT type viruses are confined to the African continent and are known for their ability to establish persistent infections and for being the source of spontaneously emerging highly infective variants (Vosloo et al., 1996). Indeed it is widely known that the African buffalo (*Syncerus caffer*) is instrumental in the maintenance of the disease for the SAT serotypes where it acts as a long-term carrier lasting for up to five years. One report has shown that in a small isolated free-living herd the disease was maintained for at least 24 years and through several generations (Condy et al., 1985). Relaxed selective constraints in these serotypes at structurally exposed proteins may permit the emergence of new mutants that may eventually enable the appearance of immune-escaping viruses becoming rapidly fixed in the population. Even though generation of immune escaping mutants is also important in the other serotypes, which is reflected also by the faster accumulation of mutations in structural proteins, the more relaxed constraints in SAT serotypes may be related to its high persistence in comparison with the other serotypes. This is supported by the magnified difference between the rates of evolution of structural and non-structural proteins when comparing SATs to non-SAT serotypes. In conclusion, genetic drift in SATs may have been an important factor in the generation of polymorphic immune-escaping mutants.

4.5.3 High rates of fixation of SDMs at SAT serotypes

Deleterious mutations are those that are eliminated by selection because of their harmful effect not only on a protein's function but on the organism fitness. These mutations are removed from the population by selection regardless of the population dynamics since the organism is deemed non-functional. If mutations are slightly deleterious (SDMs), these can be fixed in the population by genetic drift when the effective population size is small. Viral populations generally present significant sizes within their hosts but they undergo strong bottlenecks during the infection of other host individuals or in the switch to different stages during the viral life cycle.

This dynamic favours the fixation of slightly deleterious mutations. Consequently, analyses of adaptive evolution may be misleading since SDMs can be confounded by adaptive mutations. Furthermore, RNA viruses present a highly active mutational dynamic (Domingo and Holland, 1997). The fitness consequences of such mutations can be measured generally through fitness assays, which have been used to demonstrate that RNA viruses are prone to accumulate slightly deleterious mutations (Sanjuan et al., 2004b; Montville et al., 2005). A recent study has analysed the distribution of SDMs in different proteins from 143 different RNA viruses with the authors concluding that RNA viruses are indeed prone to fix such mutations (Pybus et al., 2007). Their assumption was that the average age of non-synonymous mutations increases with their selective advantage (Pybus et al., 2007; Nielsen and Weinreich, 1999). Due to that constraints at synonymous sites may inflate ω values at internal branches we used a different approach that does not rely on the assumption of neutrality at synonymous nucleotide sites.

The higher rate of fixation of SDMs in structural compared to non-structural proteins as tested in our study lends support to the greater permissibility of structural proteins to accumulate mutations by genetic drift. This pinpoints the idea that the higher diversity of structural proteins compared to non-structural ones can in part be due to the neutral fixation of SDMs. Interestingly, 1A presented no evidence for accumulation of SDMs. 1A is the only capsid protein that presents no solvent accessible amino acid regions, which may impose a higher constraint over the permissibility of amino acid sites to accumulate disruptive mutations. Accumulation of SDMs in capsid-exposed proteins may have affected the immune-escape dynamics of FMDV, which would explain the ability of SAT serotypes to establish persistent infections (Vosloo et al., 1996) much more efficiently than any of the non-SAT serotypes. This is in agreement with experimental results showing that persistent infections of African buffaloes and the genomic and antigenic diversity of SAT serotypes are positively correlated (Vosloo et al., 1996). However, we also found SDMs in proteins 3C and 3D, which would suggest also a genetic drift effect enabling the fixation of SDMs in all the proteins of FMDV. Because a greater effect of genetic drift would relax constraints in synonymous and non-synonymous sites, we should observe higher dS values at synonymous sites when genetic drift is more important. Mayrose and colleagues elegantly showed that synonymous sites are usually under constraints and that these constraints may be the cause for the inflated ω ratios in protein-coding sequences (Mayrose et al., 2007). This effect may be more dramatically enhanced by the fact that RNA viruses may be subjected to strong secondary structure constraints at nucleotide sites depending on their location in loops or stems. Our analyses of synonymous nucleotide substitutions clearly demonstrate that synonymous sites are much more relaxed in SAT serotypes than in other serotypes and significantly more relaxed in structural proteins than in non-structural proteins along the genome. The reason therefore for the increasing number of SDMs in SAT serotypes and in structural proteins may have been very likely the result of relaxed constraints in the SAT serotypes. These relaxed constraints

may have permitted the accumulation of SDMs that, once compensated for by the fixation of conditional-advantageous mutations in nearby structural regions, became advantageous to generate immune-escape mutants (explained by the increase of SDMs in structural capsid exposed proteins). To test this hypothesis we examined the phylogenetic distribution of these SDMs in each one of the serotypes as well as possible compensatory mutations being fixed in nearby structural regions on the same branches.

4.5.4 Compensatory Mutations Generate Immune-escaping Mutants in SAT Serotypes

In this study we have developed and applied a method to identify compensatory mutations (advantageous mutations conditional to their compensatory effect of SDMs). The main purpose of this test was to determine whether SDMs have accumulated as a result of either a founder effect (genetic drift), as a result of a selective process or due to a change in the selection-drift balance in each serotype. In the former case, we would expect SDMs to accumulate stochastically in terminal branches of the tree and to be located in isolated structural regions, presenting no evidence for nearby compensatory mutations. In the second case we would expect most of the SDMs accumulating in the branches leading to the serotypes (for example, they are ancestral changes that became fixed within serotypes due to its adaptive value) and present also signals of compensation in nearby structural regions. Finally in the latter case, we should observe a balanced compromise of SDMs distribution along the phylogeny and the protein structure. The fact that most CMs in structural proteins took place in the lineages leading to the South-African serotypes whereas non-structural proteins presented most of these mutations within serotypes supports that SDMs in the ancestral SAT serotypes may have been compensated to generate immune-escaping mutants establishing persistent infections. The low proportion of CMs in non-structural proteins confirms that most of mutations within serotypes are real SDMs fixed by genetic drift. We also observed a significant compensatory effect when we compared serotypes confined to Southern African to those distributed around the globe, further suggesting that SDMs combined with CMs may have allowed SAT serotypes to generate the genetic and antigenic diversity needed to escape the immune response and to establish persistent infections while maintaining their protein structural / functional stability, as previously suggested (Section 2.5.3). The results and conclusion discussed here indicate that as a consequence of its evolution, eradication of FMD from Africa as a whole is not a prospect for the foreseeable future. To complicate matters further, most countries in sub-Saharan Africa are ill- equipped to face the disease because of lack of infrastructure and financial resources. As a consequence, animal movement and migration will inevitably play a major factor in the epidemiology of the disease. The long-term result of this is despite continued efforts and extensive use of vaccines, FMD in Africa

is likely to constitute a rapidly increasing problem with more sporadic outbreaks. The danger of this is that the more developed world may be vulnerable to more undesirable transcontinental introductions of the disease due to the possibility of illegal movement of livestock or exportation of livestock products. Therefore it is imperative that adequate surveillance systems and routine sampling is maintained to monitor foot and mouth disease virus. This study has far reaching implications for the evolution and epidemiology of FMD viruses and ultimately in the control of the disease.

Chapter 5

A study of the coevolutionary patterns operating within HIV-1 using the *gag* and *env* gene

The text in this chapter partly comes from a paper published in *Molecular Biology and Evolution* (Travers, SA*, **Tully, DC***, McCormack, GP and Fares MA., 2007) in which both first and second authors contributed equally to its writing and unpublished work.

5.1 Abstract

The use of coevolution as a proxy to measure functional dependencies is evaluated here at two levels. Firstly from direct experimental evidence of the *gag* gene we examined whether or not a previously determined functional site facilitating HIV-1 adaptation in humans has imposed such a dependency between amino acid sites. If so could it be possible for a recently developed method to uncover the coevolutionary network between amino acids. Secondly we examined the *env* gene of human immunodeficiency virus (HIV). This is a functionally important gene responsible for the production of protein products (gp120 and gp41) involved in host cell recognition, binding and entry. This occurs through a complex and, as yet, not fully understood process of protein-protein interaction and within and between protein functional communication. Exposure on the surface of active HIV virions means the gp120-gp41 complexes are subjected to intense immune system pressure and have, therefore, evolved mechanisms to avoid neutralization. Using protein-coding sequences representing all of the HIV-1 group 1 subtypes we have identified amino acids within the *env* gene whose evolution is inextricably linked over the entire HIV-1 group M epidemic. We identified 848 pairs of coevolving residues, which represent 0.29% of all possible pairs. Of

the coevolving pairs 68% were significantly correlated by hydrophobicity, molecular weight or by both hydrophobicity and molecular weight. Subsequent grouping of coevolving pairs resulted in the identification of 290 groups of amino acid residues with the size of these groups ranging from two to ten amino acid residues. Many of these dependencies are correlated by function including CD4 binding, coreceptor binding, glycosylation and protein-protein interaction. In summary this analysis provides important information regarding the functional dependencies observed within HIV-1 and may assist in the identification of functional protein domains and therapeutic targets within the HIV-1 *env* gene.

5.2 Introduction

Human immunodeficiency virus type I (HIV-I) enters the host cell through a specific binding of the viral envelope glycoprotein gp120 to CD4 and a coreceptor (either CCR5 or CXCR4) (Deng et al., 1996; Dragic et al., 1996; Feng et al., 1996). The HIV *env* gene expresses a 160kD protein (gp160) that is cleaved to produce gp120 and gp41, which exist as a trimeric spike on the surface of the HIV virion containing three exterior gp120 and three gp41 transmembrane glycoproteins (Bernstein et al., 1995). The surface of the gp120 trimer is highly glycosylated with as much as 50% of the surface of the gp120 molecule being covered by carbohydrates which enables evasion of immune system recognition (Kwong et al., 1998; Wyatt et al., 1998; Pantophlet and Burton, 2006; Chen et al., 2005). A glycan shield model has been proposed whereby the gp120 glycans are continuously repositioned so as to escape neutralizing antibodies (Wei et al., 2003). It has also been proposed that the repositioning of glycans may compensate for conformational changes due to amino acid replacements occurring in virus escape from neutralizing antibodies (Pantophlet and Burton, 2006).

The binding of gp120 to the host cell CD4 receptor induces conformational changes that enable binding of a coreceptor, generally CCR5 or CXCR4, to enable host cell entry (Chen et al., 2005). Studies have suggested that, as well as the V3 loop, amino acid residues in the gp120 core around the bridging sheet are important in coreceptor binding (Rizzuto et al., 1998; Otto et al., 2003). However it is thought that the V3 loop is responsible for determining which coreceptor, CCR5 or CXCR4, is used (Hwang et al., 1991; Resch et al., 2001). It has been proposed that, following coreceptor binding, gp120 disassociates from gp41 thereby allowing access of the gp41 fusion peptide to the target cell membrane, enabling membrane fusion between the virion and the host cell membranes (Caffrey et al., 1998). Gp41 is known to comprise of four functional domains; an N-terminal fusion peptide, an ectodomain, a transmembrane domain and a cytoplasmic domain (Freed and Martin, 1995). Recently, however, it has been suggested that the C-terminal tail of gp41 may exist in two conformations, with gp41 molecules incorporated into active virions actually containing two ectodomains (termed the major and

minor ectodomains) and three membrane spanning domains (Hollier and Dimmock, 2005).

Such complexities of function and communication within and between gp120 and gp41 are reflected in the complex evolutionary patterns observed in the *env* gene (Yamaguchi-Kabata and Gojobori, 2000; Yang, 2001; Yang et al., 2003; Choisy et al., 2004; de Oliveira et al., 2004; Travers et al., 2005).

Many methods have been devised to detect selective constraints in linear multiple sequence alignments. However, intra-molecular functional relationships between amino acid sites or domains can be better understood by studying the evolutionary dependence among sites. This test in combination with other methods can yield biologically meaningful results because the dependency among amino acid sites becomes a measurable parameter when testing for coevolution (Fares, 2006). This dependence can highlight intra-protein patterns of variation used as an evolutionary strategy of the virus to escape immune response of the host and yet recognize the host cell receptor (Section 2.4.4). Studying evolution within the *env* gene using coevolution/covariation analysis has been suggested to be useful for identifying potential functional domains for mutagenesis analysis and also as selection for peptides to be used in vaccine design (Korber et al., 1993). Identifying coevolving amino acids within *env* may also aid in the identification of domains important in intra/inter protein communication as well as domains important in protein-protein interaction. While such identification of protein-protein interaction interfaces within or between proteins is theoretically possible using coevolution analyses, it is currently difficult to distinguish between the various classes of coevolving residues. The development of mathematical/statistical analytical models to distinguish between these classes of coevolving residues would be of immense benefit when testing more specific hypothesis-driven coevolutionary studies. A number of methods have been developed to detect the presence of coevolution between amino acid residues (Korber et al., 1993; Shindyalov et al., 1994; Taylor and Hatrick, 1994; Tillier and Collins, 1995; Chelvanayagam et al., 1997; Pollock and Taylor, 1997; Tuffley and Steel, 1998; Pollock et al., 1999; Pritchard et al., 2001; Galtier, 2004; Dutheil et al., 2005; Gobel et al., 1994; Lockhart et al., 1998; Tillier and Lui, 2003; Ane et al., 2005; Gloor et al., 2005). However, many of these methods are limited in that they cannot accurately distinguish phylogenetic linkage from true coevolution, they do not take into account random noise within a multiple sequence alignment or they require extremely large numbers of sequences to tackle the problem of the high rates of false positives detected. We have recently described a method that exhibits high levels of sensitivity and specificity in the detection of coevolution (Fares and Travers, 2006) and here we present the application of this method to the detection of coevolving residues within the HIV *env* gene.

In previous studies examining coevolution within HIV-1, the *env* gene have been limited to the gp120 V3 loop (Korber et al., 1993; Gilbert et al., 2005; Bickel et al., 1996). Korber and colleagues studied 308 subtype B sequences while Bickel and colleagues reanalyzed the

Korber data as well as a new dataset containing 440 sequences that represented a number of HIV-1 group M subtypes (A, C, D and E) (Bickel et al., 1996; Korber et al., 1993). Upon reanalyzing the subtype B dataset of Korber et al. (1993), Bickel and colleagues identified four of the seven coevolving pairs identified by Korber as well as a number of other coevolving pairs. However, there was no overlap of coevolving pairs identified between the analysis of the dataset containing multiple subtypes and the subtype B dataset (Bickel et al., 1996). The lack of overlap between the two sets of analyses was probably due to the use of a more relaxed strategy in the case of Bickel et al.'s study, which led to the identification of greater percentage of coevolving pairs. Gilbert and colleagues observed significant differences between the number of coevolving pairs identified in their *env* subtype B (26 pairs) and subtype C (one pair) datasets (Gilbert et al., 2005). From the results presented in these three studies it is obvious that the HIV-1 group M subtypes are exhibiting different levels of coevolution within the *env* V3 loop (Bickel et al., 1996; Korber et al., 1993; Gilbert et al., 2005). Interestingly, it is thought that the ability of the HIV strains to make the transition to CXCR4 coreceptor usage during infection may vary by subtypes. Subtype C, in particular, exhibits a lower frequency of CXCR4 usage when compared to other subtypes (Abebe et al., 1999; Ping et al., 1999; Peeters and Sharp, 2000; Cilliers et al., 2003). It is quite possible that the coevolution differences observed in the V3 loop between different subtypes represent biologically functional differences. An elegant study has been recently published that searches evolutionary convergencies (evolutionary interactions) in HIV-1 envelope (Poon et al., 2007). In this study authors applied a "covarion" like phylogenetic model to show that potential N-glycosylated sites (PNGSs) are evolutionarily linked and that exclusive interactions occur significantly more frequently between co-localised PNGSs. We have previously observed heterogeneous selective pressures operating in the evolution of the *env* gene over the HIV-1 group M subtypes (Travers et al., 2005). While the identification of subtype specific coevolution is important in identifying subtype specific evolutionary events, it is important to identify coevolving pairs/groups that are present across the entire HIV-1 group M phylogeny. The identification of such residues will provide evidence of functional, structural or interacting constraints that are conserved over the entire group M epidemic and may, therefore, identify potential functional domains for mutagenesis analysis or peptides that may potentially be used in vaccine design.

Furthermore a recent elegant study has provided for the first time compelling evidence of the adaptation of human immunodeficiency virus type 1 (HIV-1) from the cross-species transmission of simian immunodeficiency virus (SIV) infecting chimpanzees to its human hosts (Wain et al., 2007). This study demonstrated from a phylogenetic perspective and subsequent fitness assays that a single amino acid change localised to the basic N-terminal domain of the *gag*-encoded matrix (p17, MA) protein in each host species showed a marked increase in viral load in that species. The identification of the sites completely conserved among SIV infecting chimpanzees

(SIVcpz*Ptt)* strains but different in the 3 distinct groups (M, N and O) of HIV-1 did not confer an effect on fitness. However it has long been known that the evolution of amino acid sites is multi-factorial depending on the constraints imposed by their complex interaction networks (Fares, 2006). Therefore in order to evaluate the network of coevolutionary events associated with this functional important site we perform a search of the *gag* protein for evidence of functional dependencies between this reported residue and surrounding amino acids. We find a single grouping, when all HIV-1 groups are combined that is strongly correlated by important biological parameters, to be coevolving with this single important amino acid site 30 located within p17 of the *gag* encoded matrix protein. In comparison when we analyse all those available sequences belonging to SIV infecting chimpanzees (*Pan troglodytes*) we find no such evidence of an evolutionary dependency between this site and others. Our finding suggests that the strong intra-molecular network of coevolving residues acts across all groups of HIV-1 probably facilitating adaptation to its respective human hosts.

5.3 Materials and Methods

5.3.1 Taxon Selection

Firstly, protein sequences for the first section of this chapter were downloaded in the same manner as in Wain et al. (2007) and aligned using MUSCLE (Edgar, 2004). We compared sequences of HIV-1 groups M, N and O to 12 full length *gag* sequences of SIVcpz*Ptt* infecting chimpanzees (*P. t. troglodytes*) and four of the outlying but closely related SIVcpz*Pts* (from *P. t. schweinfurthii*).

Secondly, the HIV *env* gene dataset used in this study was previously described (Travers et al., 2005). For each HIV-1 group M subtype, all available full genome sequences were retrieved from the Los Alamos HIV database (`http://hiv-web.lanl.gov`) and aligned to each other using MacClade 4.08 (Maddison and Maddison, 1992). Representative sequences were selected to represent the spread of diversity throughout the subtype. This procedure was followed to avoid biased representation of the real intra-subtype diversity. Random selection of sequences from each subtype would increase the likelihood of selecting phylogenetically close sequences. For this reason sequences were carefully selected as to comprise a set of sequences spread throughout the complete evolutionary history of that subtype, based on the reconstructed phylogeny of all full genome *env* sequences for that subtype. Only a maximum of four sequences were finally selected to represent the evolutionary history of each subtype. The selected representative sequences were then manually aligned using MacClade 4.08 (Maddison and Maddison, 1992). Ambiguous regions of the alignment were removed to avoid false positives due to erroneous alignment of non-homologous sites (residues removed are as follows 6K-7Y, 12R-16R, 32E-33K,

132T-154I, 172E, 183P-190S, 310Q-311R, 320I, 354G-358T, 386N-413T, 459G-465S and 782V-788R; numbering is based on the HXB2 reference sequence). The final alignment contained 36 taxa, which represents the extent of diversity present over the entire HIV-1 group M subtypes and was 2292 nucleotides in length. A neighbour joining tree for the resulting datasets was reconstructed using PAUP* 4.0b10 (Swofford, 1998). We used this dataset to examine coevolving pairs present over the entire HIV-1 group M epidemic. In order to ensure that no biases were introduced regarding the subtype divergence levels by selecting particular sequences, we estimated the mean pairwise nucleotide divergence for each subtype in the dataset of representative sequences and compared the divergence levels between subtypes and between the representative alignment and an alignment containing all available full-genome (700) *env* sequences. Pairwise nucleotide divergences were estimated under a maximum-likelihood criterion using the model TVM + I + G which has been estimated using the program Modeltest (Posada and Crandall, 1998). The mean pairwise nucleotide distance of the full alignment of 700 sequences ($0.147 \pm 7.35 \times 10^{-5}$ nucleotide substitutions per site) and the subset used in this study for coevolution analyses ($0.155 \pm 8.0 \times 10^{-5}$) were very similar, indicating no bias in the divergence levels between both datasets. We also compared the mean pairwise nucleotide distances between the full HIV-1 dataset and the representative dataset in each one of the subtypes and the results show no significant differences (Table 5.1). To ensure that the number of sequences is not introducing any bias regarding coevolution detection we used also two datasets, with one containing all the sequences available for three subtypes (A, B and F), and the other containing the representative sequences of these subtypes. We used this approach instead of testing coevolution in the full alignment dataset due to computational limitations of the program to run over 700-sequence alignment. Finally, to discard any effect of he number of sequences in the coevolutionary analyses, we have also conducted these analyses on different subsets of the 700-based multiple sequence alignment. These subsets were built sampling randomly from each subtype the same number of sequences as in the original analyses and always those sequences showing equal divergence levels as the original set.

5.3.2 Analysis of intra- and inter-protein molecular coevolution

To test for intra- and inter-molecular coevolution, we used a recently published method for the Coevolution Analysis of Protein Sequences (CAPS) (Fares and Travers, 2006). They have previously demonstrated that the sensitivity of CAPS in detecting significant coevolving pairs is statistically significant for multiple sequence alignments containing 20 or more sequences (Fares and Travers, 2006). The method has previously been used with good effect to study coevolution within datasets containing numbers of sequences similar to those used in this study (Fares and Travers, 2006; Travers and Fares, 2007). Briefly, CAPS compares the correlated variance of the

Table 5.1: Mean nucleotide pairwise substitutions per site for the different subtypes used in this study

	Representative Data Set		All available Sequences	
Subtype	Mean	SE	Mean	SE
A	0.11684119	0.0051538	0.11765014	0.0005955
B	0.07518906	0.00590288	0.0981569	0.00014368
C	0.09584061	0.00503124	0.10160873	3.8526E-05
D	0.11271088	0.00455046	0.10666579	0.00061208
F	0.11092412	0.0053079	0.10947092	0.0033051
G	0.09416812	0.00251751	0.10147994	0.0014638
H	0.11113401	0.00836918	0.11113401	0.00836918
J	0.03589074	NA	0.03589074	NA
K	0.10280576	NA	0.10280576	NA

NOTE.- The mean nucleotide distances were estimated by maximum likelihood using the model TVM+I+G for the entire set of available HIV-1 *env* full genome sequences and for the representative set of sequences used in this study for coevolution analyses. Whenever only two sequences were available for a particular subtype, SE could not be calculated. (NA)

evolutionary rates at two sites corrected by the time since the divergence of the two sequences they belong to. This method compares the transition probability scores between two sequences at two particular sites, using the blocks substitution matrix (BLOSUM) (Henikoff and Henikoff, 1992). The significance of the CAPS correlation values was assessed by randomisation of pairs of sites in the alignment, calculation of their correlation values and comparison of the real values with the distribution of ten thousand randomly sampled values. To correct for multiple tests and for non-independence of data we implemented the step-down permutation procedure in both methods and corrected the probabilities accordingly (Westfall and Young, 1993). CAPS is implemented in the program CAPS v1.0 (Fares and McNally, 2006). For coevolution analyses, we used the protein-coding sequence, corrected for type I error using an alpha value of 0.001. The HXB2 reference sequence was used to identify the amino acid positions and all amino acid numbering presented here corresponds to HXB2. To correct for the divergence levels on each amino acid site, we weighted the correlated variability between amino acid sites by the level of substitutions per synonymous sites estimated by Li (1993).

In an attempt to identify those sites of functional importance previously detected for *gag* (Wain et al., 2007) we performed coevolution analyses on sequences from the M, N and O groups of HIV-1 together with all the SIVcpz species and then finally with just the HIV-1 sequences from all the groups. Each sub-alignment contained sufficient sequences to ensure a minimum acceptable sensitivity for the methods used (Fares and Travers, 2006).

With respect to *env* we only used full-genome representative sequences from each subtype. The selection of these sequences allowed us to definitively exclude any inter-subtype recombinant sequences that may bias the results obtained from the coevolution analyses. We have also attempted to run CAPS on the complete HIV-I sequence dataset. However, because CAPS is a

very computationally intensive method, computers could not run this program on such a large dataset. To ensure that the numbers of sequences included were not biasing the analyses, we ran CAPS on an alignment, which included subtypes A (63 sequences), B (161 sequences) and F (13 sequences). We then compared the coevolutionary results on these subtypes with those obtained when we ran CAPS on a dataset comprising the representative sequences of these three subtypes (6 sequences from subtype A, including A1 and A2, 4 sequences from subtype B and 7 sequences from subtype F, including F1 and F2). The same pairs of coevolving residues were detected in both although the correlation coefficients were slightly lower in the alignment containing the full list of sequences for the three subtypes. These coefficients were nevertheless significant at a 0.001 alpha value. The conclusion from this analysis is then that the size of the alignment does not influence the sensitivity of the coevolution analysis as far as the multiple sequence alignment contains more than 10 sequences, something already shown in a previous work (Fares and Travers, 2006).

Molecular coevolution can be divided into many different types including structural, functional, interaction, phylogenetic and stochastic coevolution (Atchley et al., 2000). Disentangling the different types of coevolution is anything but straightforward. In a previous work, however researchers attempted to distinguish between phylogenetic, stochastic and the other components of coevolution through a phylogenetic-based co-evolution analysis procedure (Fares and Travers, 2006). Distinguishing between structural, functional and interaction coevolution requires biological information in addition to the mathematical adjustments made to the method. Estimating the correlated variation in hydrophobicity, molecular weight or combination of both parameters may introduce further information regarding the coevolutionary relationships (functional, structural or functional and structural) among covarying sites. We therefore conducted an analysis of correlation between coevolving amino acid sites taking into account these biological parameters.

5.3.3 Mapping significant amino acid residues onto 3D protein structures

The three-dimensional structure of the human immunodeficiency virus type 1 matrix protein (PDB accession number 2HMX) was used to map significant amino acid residues coevolving with site 30. In addition many three-dimensional structures have been resolved for the gp120 and gp41 proteins, but we used structures representing gp120 in complex with a CD4 receptor and a neutralizing antibody (PDB accession number 1G9M), an unliganded simian gp120 core structure (2BF1), the V3 loop from a V3 loop containing gp120 structure (2B4C) and a structure representing the SIV gp41 ectodomain (1IF3) (Caffrey et al., 1998; Kwong et al., 1998; Chen et al., 2005; Huang et al., 2005). The 3D structure viewing and manipulation was performed using iMOL (`http://www.pirx.com/iMol`).

Table 5.2: Residues detected as coevolving with site 30 and correlated by hydrophobicity (Hydro) and molecular weight (MW). The denotation of the number 1 indicates significant correlation while 0 indicates no significant correlation has been detected. Probabilities (Prob) associated with each biological parameter are shown in the subsequent column. Dashed lines (-) indicate no significance.

Site 1	Site 2	Coevolution	Hydro	MW	Hydro Correlation	Prob	MW Correlation	Prob
30	46	1	1	0	0.2908	0.043	-	-
30	60	1	1	0	0.6028	0.0121	-	-
30	105	1	1	0	0.3421	0.0355	-	-
30	332	1	1	0	0.6597	0.0079	-	-
30	467	1	1	1	0.762	0.0039	0.3229	0.0433

We used a conservative mean distance of 8Å in determining whether two amino acid residues were significantly proximal in the 3D structure. The relative distance between two amino acids was calculated by taking the mean three-dimensional atomic coordinates for each amino acid in the structure. We then calculated the distance between two amino acids as the distance between their mean coordinates as follows:

$d = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$

where x, y and z are the mean atomic coordinates for residue 1 and x', y' and z' are the mean atomic coordinates for residue 2.

5.4 Results

5.4.1 Coevolution of a functionally important residue in *gag*

Using the same data as in a previous study (Wain et al., 2007), our intra-molecular coevolution analysis of the *gag* gene of HIV-1 groups M, N and O reveals a single grouping comprising six amino acids all of which have a significant dependency with site 30K (Table 5.2 and Figure 5.1). Five of these seven are located within the encoded matrix protein and one localised to the capsid protein and the small p6 region. Interestingly when SIVcpz strains were analysed at alpha values of 0.05 and 0.001 respectively no single grouping containing this site was observed. Similarly when all the HIV-1 groups (M, N and O) were analysed with SIVcpz sequences this grouping disappeared. However, upon analysing each HIV-1 group separately with SIVcpz at least a single grouping with site 30 was reported. This indicates that the detection of this site to be coevolving is as a direct result of the evolution of HIV-1 and SIVcpz. Furthermore, this single grouping detected to be functionally coevolving is highly correlated with two relevant biological parameters - molecular weight and hydrophobicity. These two are the most important parameters in explaining functional links between coevolving amino acids.



Figure 5.1: 3D structure of the p17 *gag* encoded matrix protein for HIV-1. This figure indicates those sites coevolving with the known functional important site previously detected. Space filled residues coloured in blue signify those sites detected to be coevolving with site 30 in this domain. The space filled amino acid residue in yellow indicates the single functional important site required for adaptation.

5.4.2 Coevolution analysis in *env* gene belonging to group M

Coevolution analysis resulted in the identification of 848 pairs of coevolving residues representing 0.29% of all possible pairs. The observation of coevolving pairs constituting 0.29% of all possible pairs in this study is significantly lower than those previously reported for HIV-1. For example 1.33% by Korber et al. (1993), 5.24% observed by Gilbert et al. (2005) for their subtype B data set and 12.69% observed by Bickel et al. (1996) in both their 308 and 440 data sets. The mean correlation coefficient for these pairs was 0.5962 (range 0.5000-0.9944). These pairs represented 233 amino acid residues within the *env* gene, 145 and 88 within gp120 and gp41 respectively. Subsequent grouping of all pairs resulted in 290 groups of coevolving residues with the size of these groups ranging from two to ten amino acid residues. The majority of these groups (72%), however, contained either two or three residues (Table 1 available at `http://mbe.oxfordjournals.org/cgi/content/full/msm213/DC1`).

We also applied further filters to identify coevolving pairs whose coevolution was correlated by hydrophobicity, molecular weight or both (Table 5.3). This analysis would enable identification of compensatory mutations and/or mutations at structurally related amino acid sites. Of the 848 pairs of coevolving residues (involving about 263 out of 764 amino acid sites) identified in *env*, 311 and 268 of these were correlated by hydrophobicity and molecular weight respectively, whereas 194 residues were correlated by both hydrophobicity and molecular weight. The 848 coevolving pairs together with their co-evolutionary parameters are available from `http://mbe.oxfordjournals.org/cgi/content/full/msm213/DC1`

In order to visualize the spread of coevolving pairs throughout the *env* gene we plotted a matrix exhibiting coevolving pairs and also pairs whose coevolution was correlated by hydrophobicity, molecular weight or both hydrophobicity and molecular weight (Figure 5.2). Although coevolving pairs were spread throughout *env*, two distinct regions exhibit high levels of coevolution with many residues in *env*; the end of C2 with V3 and C3 as well as a portion of the gp41 cytoplasmic domain (Figure 5.2).

5.4.2.1 Coevolution within the gp120 V3 loop and proposed co-receptor binding domains

Previous coevolution studies in *env* focused on identifying coevolution within the V3 loop (Korber et al., 1993; Bickel et al., 1996; Gilbert et al., 2005). We have expanded upon these studies and investigated the presence of coevolution throughout the entire *env* gene. Within the V3 loop, however, we have identified 24 pairs of coevolving residues comprising 14 residues of the V3 loop (Figure 5.3). Of the 24 pairs of coevolving residues, four (17%) of these were significantly

Table 5.3: Details of the coevolving pairs whose coevolution was correlated by hydrophobicity, molecular weight, or both.
NOTE.- The number of significant pairs is shown as are the mean values and ranges for the correlation and probability statistics. Also shown is the number of correlated residues located within the *env* gp120 and gp41 domains.

	Number of Significant pairs	Mean Correlation	Mean Probability	gp120	gp41
Hydrophobicity	311	0.459403 [-0.1660 - 0.9877]	0.0182053 [0.0015 – 0.0500]	123	69
Molecular weight	268	0.396898 [0.1653 – 1.000]	0.0217441 [0.0015 – 0.0494]	113	64
Hydrophobicity and	194	Hydro: 0.525354 [0.1511 – 0.9877]	0.014571 [0.0015 – 0.0482]	95	56
molecular weight		Mw: 0.431891 [0.1653 – 1.000]	0.0190494 [0.0015 – 0.0494]		

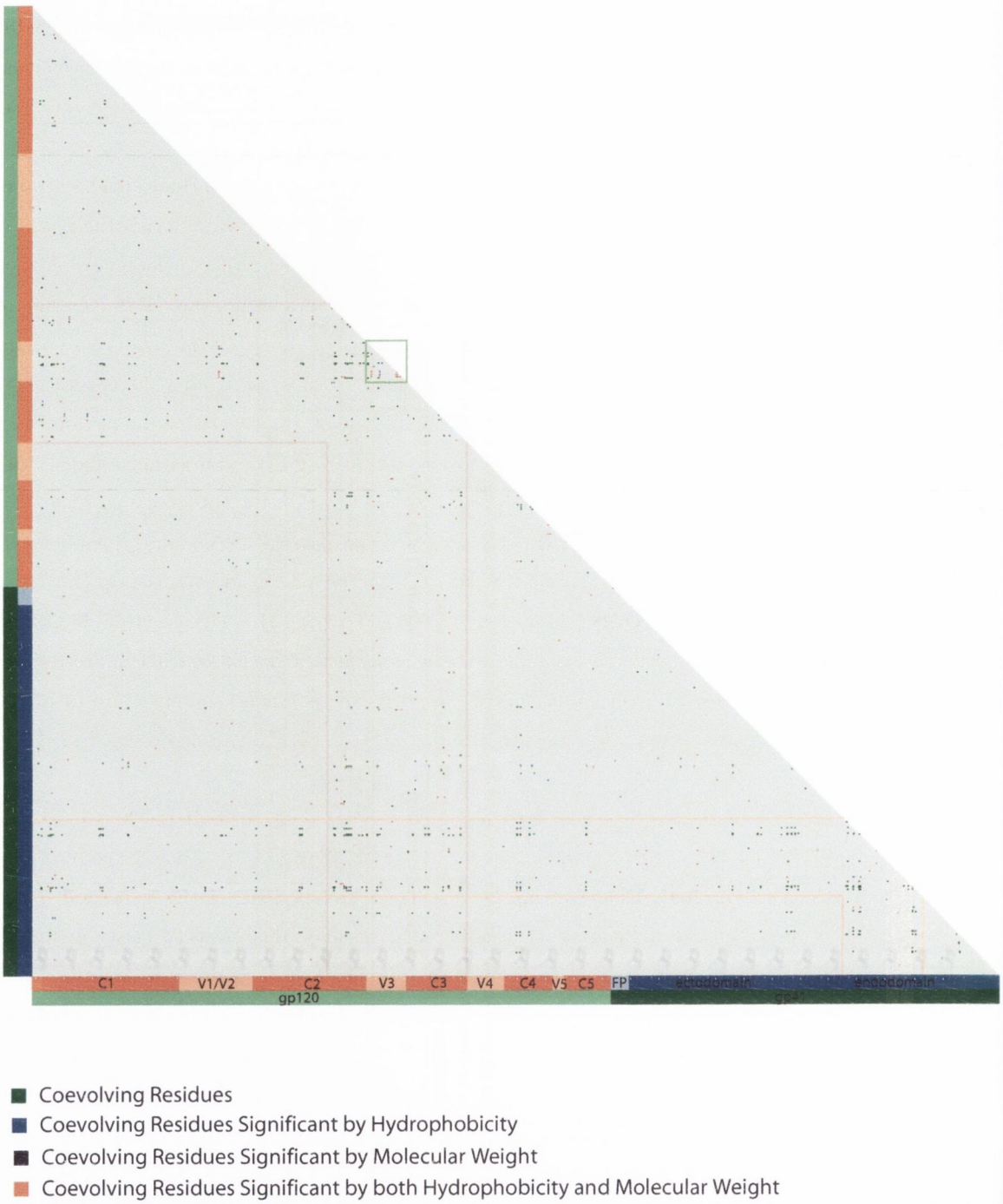


Figure 5.2: Coevolving pairs of amino acid residues over the complete *env* gene. Pairs whose coevolution is correlated by hydrophobicity, molecular weight or both hydrophobicity and molecular weight are marked. Also shown between coloured lines are the two regions that appear to exhibit higher levels of coevolution than that observed over the entire *env* gene. Coevolving residues within the V3 loop are also shown (green box).

correlated by hydrophobicity, while 13 (54%) were significantly correlated by both hydrophobicity and molecular weight. Five of these 13 pairs of coevolving residues had also been identified as coevolving by Bickel et al. (1996) in their 440 data set, which contained representative sequences from multiple subtypes.

Upon solving the structure of a V3-containing HIV-1 gp120 core, Huang et al. (2005) proposed that, following CD4 binding, the N-terminus of the CCR5 receptor binds the gp120 core and V3 base while the V3 tip binds the co-receptor's second extracellular loop. Complimentary to this, Rizzuto et al. (1998) identified a number of residues within the gp120 core, the mutation of which significantly affects CCR5 co-receptor binding by gp120. Many of these residues were not identified as coevolving within *env*, most likely because of functional conservation. Rizzuto and colleagues identified two residues, P437 and Q442, that, when mutated, result in a $\geq$50% increase in CCR5 binding with respect to wild-type gp120 (Rizzuto et al., 1998). Both of these residues were identified as coevolving with residues within the V3 loop stem (N302 with P437 and R306 with Q442). Also within the proposed gp120 core CCR5 binding domain residue V200 was identified as coevolving with three residues in the V3 loop, two in the tip (R315 and A316) and one in the base (Q328). Mutation of V200 showed a slight decrease (16%) in CCR5 binding when compared with wild type (Rizzuto et al., 1998). Residue G379, while not tested by Rizzuto and colleagues, is directly adjacent in the gp120 three-dimensional structure to E381 (5.38Å) which, when mutated, decreases CCR5 binding by 93%. G379 was observed in this study as coevolving with T297 in the base of the V3 loop.

CD4 binding has been shown to induce conformational changes in gp120 (Chen et al., 2005), which have been proposed to expose domains within gp120 responsible for co-receptor binding (Chen et al., 2005; Huang et al., 2005). Therefore, because of the structural dependencies between the CD4 and co-receptor binding domains one would expect to see a certain degree of coevolution between these domains or their neighbor peptide regions to maintain function. We have observed eight residues within the V3 loop and five residues within the proposed co-receptor binding domain on the gp120 core that coevolve with residues that either bind directly to CD4 or are proximally contained within the CD4 binding pocket ($<$8Å from residues that bind CD4 directly, table 5.4.2.1).

5.4.2.2 Networks of coevolving amino acids between gp120 and CD4

Core to the function of HIV is the binding of gp120 to the CD4 receptor on the host cell surface. We have investigated the coevolution network present within amino acid residues involved in CD4 binding by gp120. Included in this were residues that bind directly to CD4 (Kwong et al., 1998), residues that comprise the epitope for BMS-806, the binding of which interferes with gp120-CD4 binding (Pantophlet and Burton, 2006) as well as other residues contained within

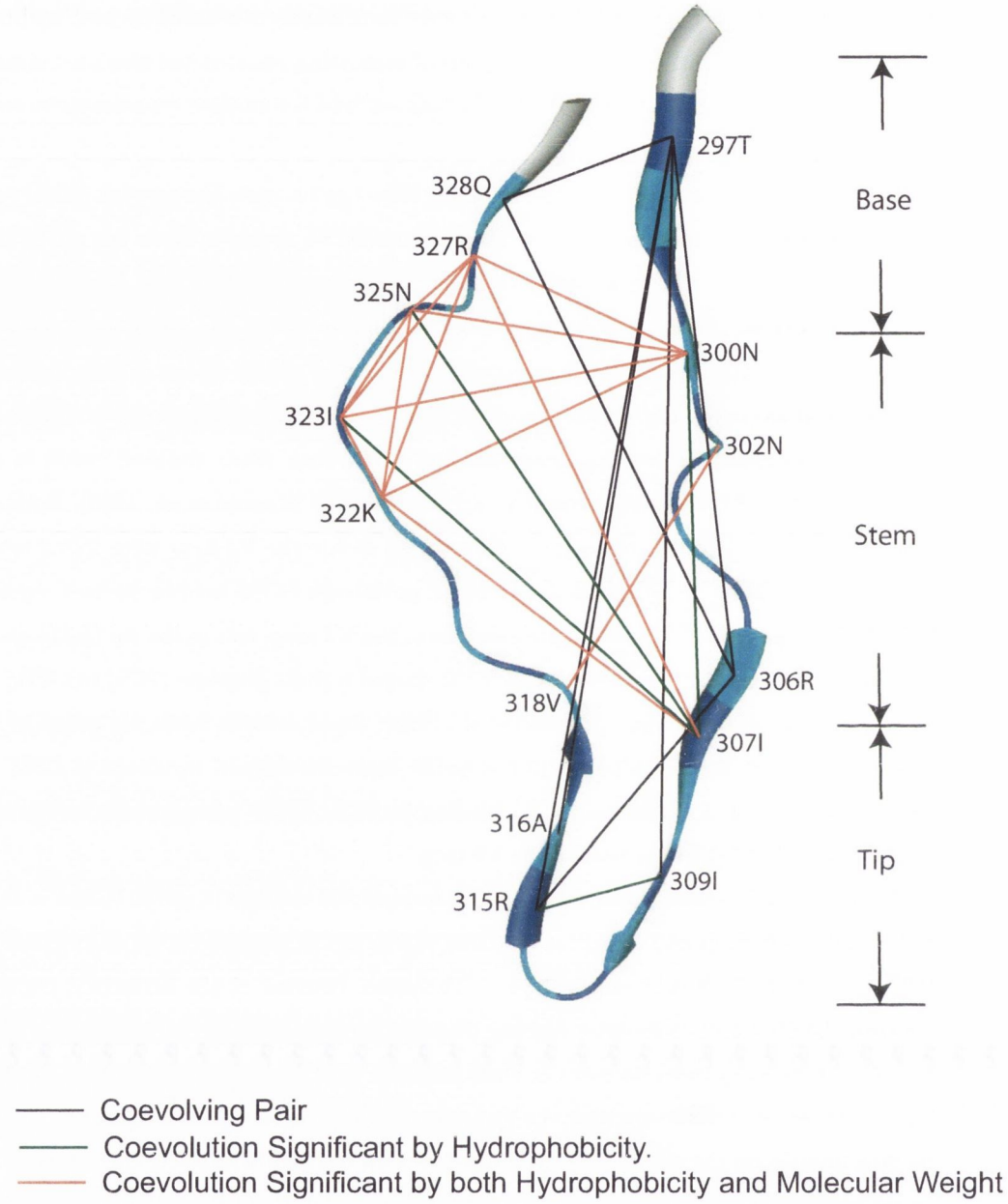


Figure 5.3: Pairwise coevolving residues observed within the gp120 V3 loop (Huang et al., 2005). Numbering is per the HXB2 reference sequence, and residues are colored alternately to ease visualization.

Table 5.4: Residues in the V3 Loop and proposed CCR5-Binding domain that coevolve with residues involved in CD4 binding.

	Bind directly to CD4	Directly adjacent to CD4-binding pocket
V3 loop		
T297	T283, K429	S364 (S365, 4.36451Å)
R306	-	S364 (S365, 4.36451Å)
I307	-	T278 (D279, 5.34946Å)
I309	T283	-
R315	D279, A281, T283	-
A316	D279	-
R327	-	S274 (T283, 6.72668Å), N276 (D279, 5.35118Å)
Q328	D279, A281, T283	-
CCR5-binding domain		
K121	S365	-
P437	-	T373 (I371, 6.34689Å)
R440	D279, K429	-
Q442	D279	S364 (S365, 4.36451Å), K432 (N425, 5.96993Å)
R444	-	K432 (N425, 5.96993Å)

NOTE.- Also shown in brackets are the closest adjacent CD-4 binding residues and the pairwise distance in angstroms (Å).

conserved CD4 binding site epitopes detailed by Wyatt and colleagues (Wyatt et al., 1998). We have also included amino acid residues, which may be functionally proximal (<8Å) to residues responsible in CD4 binding in both the HIV liganded and SIV unliganded gp120 structures (Kwong et al., 1998; Chen et al., 2005). The CD4 coevolution network contained 32 amino acid residues (Figure 5.4), eight of which bind CD4 directly, one that maps to the BMS-806 epitope and five residues that are directly glycosylated in the HIV or SIV structures (these are located <8Å from residues important in CD4 binding). Of the 37 coevolving pairs present in the CD4 network, the coevolution of 65% of these was correlated by hydrophobicity (five pairs), molecular weight (two pairs) or both hydrophobicity and molecular weight (17 pairs).

The "glycan shield" model suggests that domains within the gp120 structure are protected from neutralizing antibodies by the presence of carbohydrate molecules bound to the surface (Wei et al., 2003). This shield is formed within the gp120 tertiary structure bringing linearly distant domains into close proximity to form the shield. Therefore one would expect that, in order to maintain the overall structure of the glycan shield, there would be a degree of coevolution between directly glycosylated residues and also residues directly proximal to glycosylated residues (as mutation at these residues could affect the overall structure of the shield). This is, in fact, the case with an extensive network of coevolution observed between 41 directly glycosylated and glycosylation-related residues (Figure 5.5). The coevolution of 67% of the 42 coevolving pairs

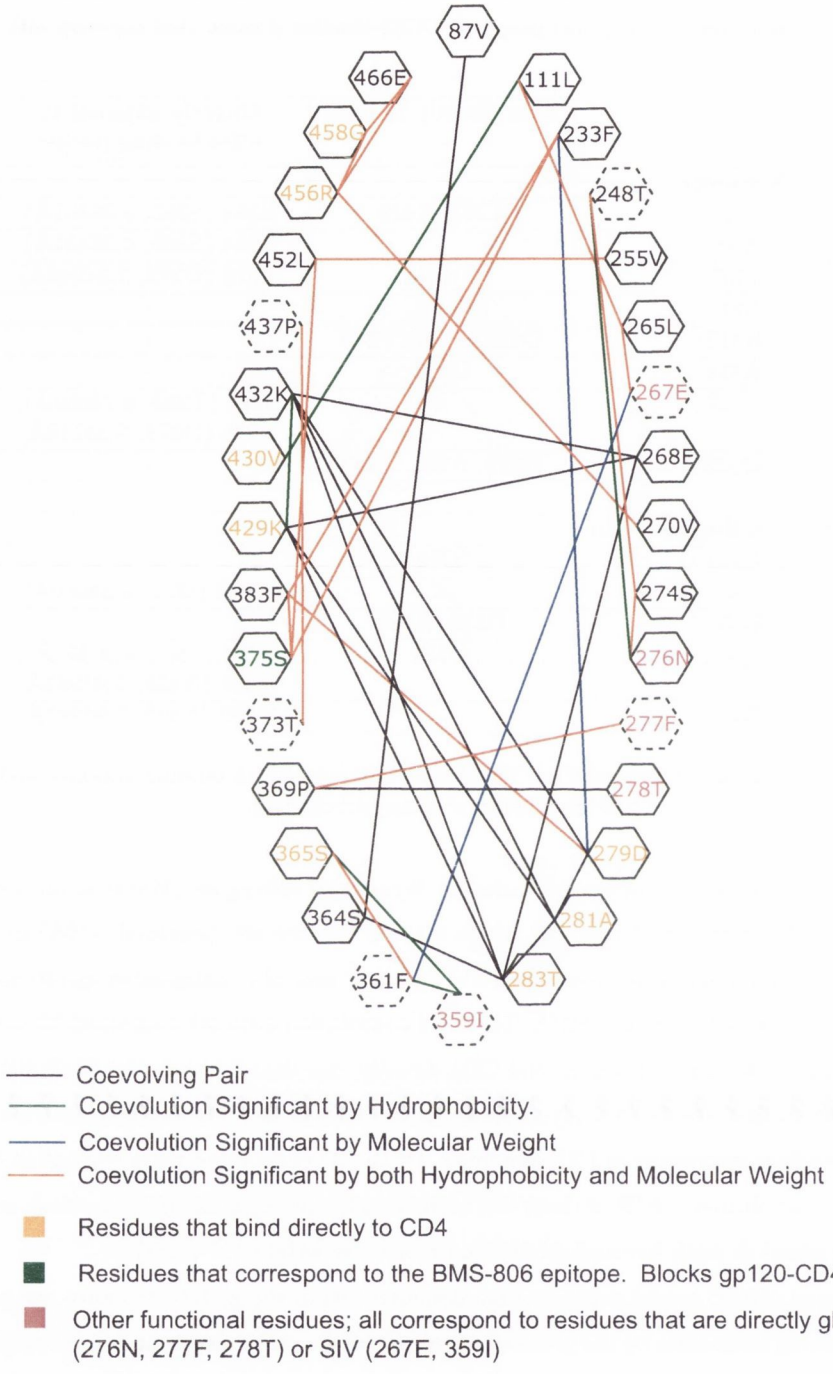
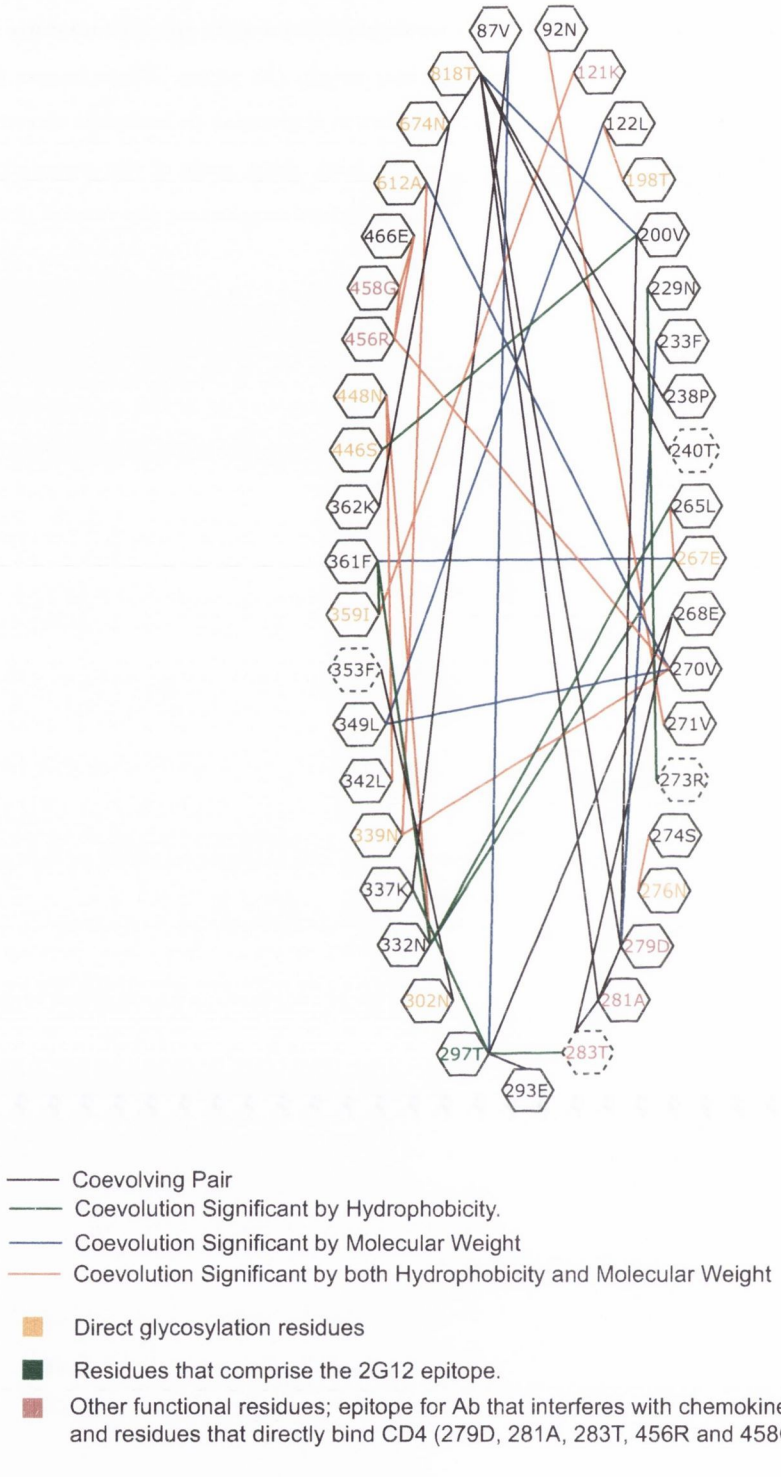


Figure 5.4: The CD4 coevolution network. Included are residues that bind CD4 directly and residues that correspond to the BMS-806 epitope, the binding of which interferes with gp120-CD4 interaction. All other residues are within 8Å of CD4 functional residues. A number of these proximal residues correspond to known glycosylation residues and these are also marked. Proximal residues in dashed pentagons were only observed as proximal in the SIV unliganded gp120 structure. All other residues are proximal in both the HIV liganded and SIV unliganded gp120 structures.

in the glycosylation network was correlated by hydrophobicity (eight pairs), molecular weight (seven pairs) or by both hydrophobicity and molecular weight (13 pairs). These results support that coevolution between or nearby N-glycosylated sites is important to maintain the structure of the glycosyl shield against the defense system of the host. Also, most of the coevolving pairs included sites that were not directly proximal in the structure supporting the results previously reported (Poon et al., 2007).



Coevolving Pair
Coevolution Significant by Hydrophobicity.
Coevolution Significant by Molecular Weight
Coevolution Significant by both Hydrophobicity and Molecular Weight

Direct glycosylation residues
Residues that comprise the 2G12 epitope.
Other functional residues; epitope for Ab that interferes with chemokine receptor binding (121K) and residues that directly bind CD4 (279D, 281A, 283T, 456R and 458G)

Figure 5.5: The glycosylation coevolution network. Residues that are directly glycosylated are shown, as are residues that comprise the 2G12 epitope. Similarly to the CD4 network, all other residues are within 8Å of glycosylation functional residues that are directly glycosylated or correspond to the 2G12 antibody epitope. Proximal residues in dashed pentagons were only observed as proximal either in the HIV liganded structure (240T and 353F) or in the SIV unliganded gp120 structure (273R and 283T).

We also observed a large degree of overlap between the CD4 and glycosylation networks with 63 pairs of coevolving residues observed between these. Of the 63 coevolving pairs, 12 were present in both networks, 42 were present in one of the networks while nine novel pairs had not been identified in either the CD4 or glycosylation coevolution networks.

5.4.2.3 Coevolution within gp120 moving domains

Recently the structure of an unliganded SIV gp120 core was resolved (Chen et al., 2005) and showed marked differences with a structure of a gp120 core liganded with CD4 (Kwong et al., 1998). Chen and colleagues observed large displacements within the gp120 core inner domain and the absence of the bridging sheet in the unliganded structure. With the exception of two regions, the orientation of the outer domain remained essentially the same between the two structures. Using both the liganded and unliganded gp120 structures we looked for coevolving residues that showed a significant difference (>8Å difference) in their mean pairwise distances between the two structures. We identified six coevolving pairs that were significantly more proximal in CD4 bound gp120 than in unliganded gp120 and three coevolving pairs that were significantly closer in the unliganded gp120 structure (Table 5.4.2.4). With the exception of two pairs (P369&G379 and E102&S364), only one of the coevolving residues in each pair is located within a domain identified by Chen et al. (2005) as moving significantly following CD4 binding.

5.4.2.4 Inter gp120-gp41 coevolution

The three-dimensional structure of the ectodomain of SIV gp41 has been solved and combining the properties of this structure with previous mutagenesis analyses Caffrey and colleagues proposed that the gp120-binding domain is located within a hydrophobic patch within the gp41 loop domain (Caffrey et al., 1998). We observed four residues that map to this hydrophobic patch (K588, T605, A607 and A612) as coevolving with residues elsewhere within both gp120 and gp41 (Table 5.6, Figure 5.6). Residue K588 coevolves with three residues (M535, Q543 and H564), all of which map to the FP/NHR domain within the gp41 ectodomain (Figure 5.6A) , which is proposed to move out from gp120 and enable fusion of the virion and target cell membranes (Caffrey et al., 1998). The only other residue in the gp41 ectodomain that K588 coevolves with is E662 that, while located on the CHR, appears to be located in the corresponding position on the CHR as Q543 is on the NHR (Figure 5.6A).

Table 5.5: Coevolving amino acid residues that exhibit marked differences in their mean pairwise distances between the liganded HIV gp120 and the unliganded SIV gp120 structures.

Coevolving Pair	Mean pairwise distance in liganded gp120 (Å)	Mean pairwise distance in liganded gp120 (Å)
Coevolving residues proximal in liganded gp120		
92N[a]and 271V[b]	19.0935	28.7078
111L[a] and 265L[b]	24.1784	32.5983
122L[a] and 471G[b]	24.5348	34.9453
240T[a] and 348K	22.4094	31.5596
369P[a] and 379G[a]	19.1430	32.4727
369P[a] and 440S[c]	23.4603	35.4941
Coevolving residues proximal in unliganded gp120		
87V and 106E[a]	31.0536	22.5208
87V and 364S[a]	43.9368	29.8071
102E[a] and 364S[a]	26.6992	16.1491

NOTE.- Distances are shown in angstroms (Å).

[a] Residues which exhibit movement between the liganded HIV and unliganded SIV gp120 structures

[b] Pairs whose coevolution is correlated by both hydrophobicity and molecular weight

[c] Pairs whose coevolution is correlated by molecular weight

Table 5.6: Residues within the proposed gp120 binding hydrophobic patch of the gp41 ectodomain and residues elsewhere in *env* that they coevolve with.

gp41 hydrophobic patch residues	Coevolving residues
588K	17W[a], 277F[b], 369P, 373T[a], 535M[b], 543Q[c], 564H[a], 662E[a], 726G, 746I[a], 758D[a], 777I[a]
605T	270[c], 339N[c], 683[a]
607A	778V
612A	270V[b], 339N[a]

[a] Pairs whose coevolution correlates by both hydrophobicity and molecular weight
[b] Pairs whose coevolution correlates by molecular weight
[c] Pairs whose coevolution correlates by hydrophobicity

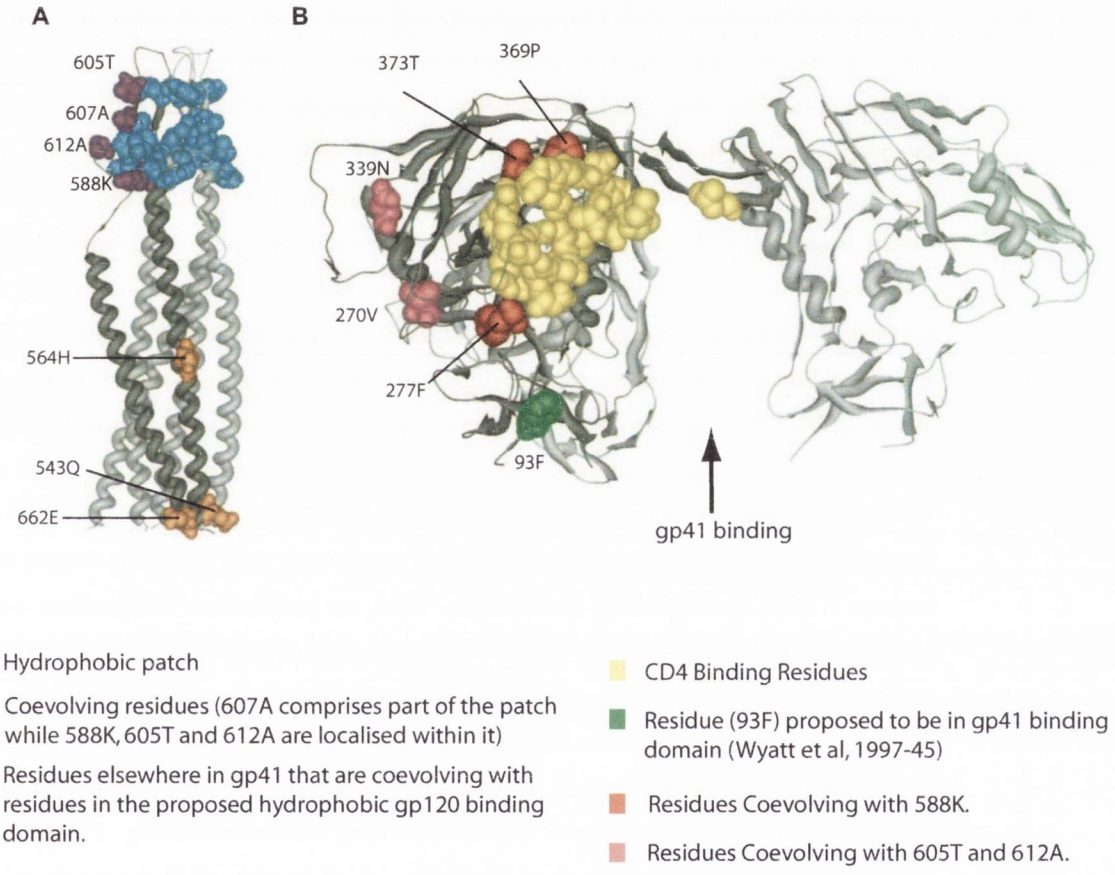


Figure 5.6: Inter-protein coevolution between gp41 and gp120. (A) Coevolving residues within the gp41 ectodomain and (B) the residues within the gp120 core with which they coevolve. NB: With the exception of the hydrophobic patch all marked residues are only shown on one molecule In each trimer, the labeled molecules are coloured darker.

Both T605 and A612 coevolve with V270 and N339 which map to the highly glycosylated outer domain of the gp120 core structure, while K588 coevolves with F277, P369 and T373 all of which map to the CD4 binding domain in gp120 (Figure 5.6B). With the exception of W17 and K683 the remaining five residues that coevolve with the gp41 hydrophobic domain (G726,

I746, D758, I777 and V778) are located within the gp41 cytoplasmic domain.

5.5 Discussion

The application of this robust coevolution method has lead to the detection of a known amino acid of functional importance (Wain et al., 2007). This indicates that the method is performing well and hence we further applied it to the *env* gene of sequences from the global HIV-1 pandemic.

5.5.1 Co-adaptation of *gag* to its human host

As amino acids seldomly evolve independently they leave a footprint defining their intra-molecular functional relationship. This relationship can be better understood by studying the evolutionary dependence among sites. Such evolutionary interdependency between protein residues creates frequent functional and structural communication between sites. The case of Wain et al. (2007) provides a good test to evaluate whether or not a functionally important site known to facilitate the adaptation of HIV-1 into human hosts has such dependency. They revealed a single site located within the *gag* polyprotein that enhances the replication potential of primary lentiviruses following transmission to humans. From our results we preclude the independence of the single site 30 within p17, the gag-encoded matrix, and reveal through the use of intra-molecular coevolution that this site has a strong dependency between other amino acids in p17 and in the more conserved domain of p24 also referred to as the capsid protein and the small p6 domain. We firmly believe that the non-conservative amino acid replacement at position 30 on three independent branches leading to the three HIV-1 groups may have created a compensatory effect between a suite of surrounding amino acids. Thus, our results show independent dependencies between site 30 and other protein residues that exist on all the branches leading to the three HIV-1 groups. The location of these coevolving residues is also of interest as the first 31 amino acids of p17 have been reported to be a membrane targeting domain that mediates binding to acidic membrane phospholipids (Zhou et al., 1994). It is now believed that interactions between this membrane and these phospholipids may enhance Gag membrane binding. Further, the capsid-encoded protein of gag has an important role in virus assembly and is also involved in early cell entry. Similarly p6 has a major function to promote the detachment of assembled virions from the cell surface and for each other. Therefore the detection of a grouping of coevolving residues within these functional domains indicates the wide level of interactions taking place within gag. It highlights the strong functional/structural dependency between protein residues and suggests an intrinsic network of coordinated interactions between domains is at play. In conclusion, we demonstrate the complexity of the evolutionary dependency between a non-conserved functional site that mediates efficient replication in the human host. Such functional communication is

further exacerbated from examination of the three-dimensional structure that implies that due to the coevolving sites not being proximal (Figure 5.1), their relationship may be reflected in the functional equilibrium of important regions within the protein. As such the growth advantage conferred with a substitution at site 30 in HIV-1 groups may have in turn be as a direct result of a network of interactions between important functional domains thereby facilitating the species jump of HIV from chimpanzees to humans.

5.5.2 Coevolutionary patterns operating within *env*

A further study following on from the first (above) is the evaluation of coevolution operating within the *env* gene across all the HIV-1 group M subtypes. The identification of pairs or groups of coevolving residues provides a wealth of information with regard to amino acid residues or protein domains that exhibit dependency in their evolution. We have, where possible, connected the coevolution results to biological knowledge. The remaining coevolution pairs/groups presented here (Table 1 available at `http://mbe.oxfordjournals.org/cgi/content/full/msm213/DC1`) should be viewed as potentially biologically significant pairings and we suggest that many of these results should be further examined experimentally to determine the biological significance of the observed coevolution. We must emphasize, however, that the observation of coevolution between two domains does not indicate protein-protein interaction between these domains. We have suggested other reasons for the observation of coevolution both here and in previous works (Fares and Travers, 2006; Travers and Fares, 2007). Representative sequences were selected in such a way as to sample a complete cross section of the diversity observed within each subtype (Travers et al., 2005). Although still not fully understood, the functional complexities of the gp120-gp41 complex have been well documented (Wyatt et al., 1997; Poignard et al., 2001; Chen et al., 2005; Hartley et al., 2005; Pantophlet and Burton, 2006; Kwong et al., 2000). Coupled with this, the intense selective pressures known to operate on *env* have resulted in a gene with incredibly complex, multi-faceted evolutionary dynamics (Holmes et al., 1992; Seibert et al., 1995; Yang, 2001; Choisy et al., 2004; de Oliveira et al., 2004; Travers et al., 2005). In this study we have attempted to improve the understanding of the coevolutionary selective pressures operating on the *env* gene across all the HIV-1 group M subtypes. While previous studies have concentrated on examining coevolution within the gp120 V3 loop (Bickel et al., 1996; Gilbert et al., 2005; Korber et al., 1993) and others have observed subtype specific patterns of evolution (Gnanakaran et al., 2007; Korber et al., 1994; Gaschen et al., 2002) the advent of more sensitive and accurate methods has enabled us to examine coevolution across the entire *env* gene. The use of a highly sensitive method such as CAPS has allowed us to identify a significantly smaller subset of coevolving pairs in HIV-1 *env* gene compared with previous works. The sensitivity of CAPS has been estimated to be around 95% in multiple sequence alignments comprising around

40 sequences (Fares and Travers, 2006). Our confidence on the low proportion of false positives based on previous works allows us to confirm that most of the amino acid site pairs detected are true positive coevolving pairs. This work has been also applied in other case studies showing the high accuracy of the method in detecting true positive results (Fares and Travers, 2006). In contrast to previous studies, we have studied a data set containing representative sequences from all HIV-1 group M subtypes as opposed to single subtypes or a number of subtypes together. The identification of such coevolving residues can provide insights into domains of proteins or pairs of amino residues within a protein whose evolution is inextricably linked by structural, functional or interacting constraints. In this study, 46% of all coevolving pairs were correlated by hydrophobicity, molecular weight or by both hydrophobicity and molecular weight. Some of these correlated pairs are linearly proximal, for example, 456R and 458G. However, some of these correlated pairs are linearly distant but structurally proximal, for example, 122L and 198T are separated by 76 amino acids on a linear level yet are only 6.7Å apart in the HIV gp120 3D structure (Kwong et al., 1998). Such correlations of coevolving pairs are testament to the evolutionary complexities operating within HIV. The absence, however, of complete structural data and deeper comprehension on the mode of virus operation makes the distinction of the type of amino acid sites dependency anything but straightforward. We cannot exclude the effect of recombination in our results of coevolution. Even though the selection of representative full-genome sequences of each subtype allowed us to avoid the effects of inter-subtype recombination, excluding intra-subtype recombination remains a problem. However, currently there is no way to identify intra-subtype recombination and, as with all analyses with HIV-1 group M multiple sequence alignments, therefore some intra-subtype recombinants may be present in the data. Discussion of the biological significance of the complete set of coevolving pairs (848 pairs) is impossible in this manuscript because there is no reported functional data on each one of the pairs. Also, only simulation studies can provide a measure of the sensitivity, and thus of the amount of positive results, of the method to detect real coevolution. Several lines of evidence indicate that these pairs are not false positive resulting from a limited statistical power of the method used. First comparison of the correlation coefficients of each non-discussed pair of coevolving sites with that for the pairs with biological information show no difference in their values. Second, our previous analysis of the performance of the method to detect coevolution (Fares and Travers, 2006), using a simulation approach developed in other works and not related to our algorithm to detect coevolution, showed that the sensitivity of the method can be as high as 90% when the number of sequences in the multiple sequence alignment is above 20, although we have used 36 sequences in our study. Despite this fact, we still believe that a minor fraction may be false positives, although this fraction is dramatically smaller than in other studies performed so far.

5.5.3 The distribution of coevolving residues throughout the *env* gene

While coevolving residues are spread throughout the *env* gene, we did observe two regions that present a higher density of coevolution with residues throughout the *env* gene (Figure 5.2). The first of these regions covers the latter part of C2 as well as the V3 loop and C3 domain. Observing such a density of coevolution in this region is not at all surprising as it contains a large proportion of amino acid residues involved in glycosylation as well as CD4 and chemokine receptor binding (Kwong et al., 1998, 2000; Wei et al., 2003; Rizzuto et al., 1998; Wyatt et al., 1998; Chen et al., 2005). Coevolution between residues within this domain is most likely occurring to maintain the overall structural properties required for optimum protein function. The binding of gp120 to CD4 is known to induce conformational changes within gp120 (Chen et al., 2005), which have been proposed to make further gp120 domains accessible for coreceptor binding (Chen et al., 2005; Hartley et al., 2005; Huang et al., 2005). Following coreceptor binding, it has been proposed that gp120 disassociates from gp41 to enable gp41-facilitated cell membrane fusion (Caffrey et al., 1998). This complex mechanism requires a large degree of intra- and inter-domain communication within and between the gp120 and gp41 molecules. The large degree of coevolution present between the C2-V3-C3 region and residues throughout *env* supports this claim.

The second region exhibiting a high level of coevolution with residues throughout *env* is interesting as it is located within the gp41 cytoplasmic domain. Hollier and Dimmock (2005) detailed a number of studies which have shown that antibodies specific to an antigenically active motif (Kennedy sequence) in the gp41 cytoplasmic domain can neutralize HIV-1 virions (Hollier and Dimmock, 2005). As antibodies cannot cross the lipid bilayer of the cell membrane this suggests that a portion of the gp41 cytoplasmic domain is exposed on the cell surface. Based on this evidence Hollier and Dimmock (2005) proposed a structural model for gp41 that consists of three membrane spanning domains (MSDs) and two ectodomains, a major and a minor. The Kennedy sequence is exposed on the outer face of the proposed minor ectodomain. Hollier and Dimmock (2005) suggested that this gp41 structure is evident only in a minority of cell-associated gp41 molecules that are destined for incorporation into active virions. Of the nine residues within gp41 that show high levels of coevolution with residues elsewhere in *env*, five of these (L721, G726, E731, G732 and I746) are located in the minor ectodomain, with three of them (G726, E731 and G732) being located within the Kennedy sequence. It has been suggested that there may be interactions between the minor ectodomain and the major ectodomain as well as with elements of gp120 and also with other gp41 monomers that form the gp41 trimer (Hollier and Dimmock, 2005). This level of functional dependency among domains within the gp120-gp41 complex would explain the large degree of coevolution observed between the five residues located within the minor ectodomain and residues elsewhere in *env*. The remaining four residues that

comprise the region of gp41 coevolving with a large number of residues throughout *env* (L774, V778, T779 and I781) are located within the cytoplasmic domain of gp41. All of these residues are directly adjacent to the second tyrosine-dependent sorting signal in gp41 ^{768}YHRL^{771} (Hollier and Dimmock, 2005), a peptide of which has been shown to interact with an adaptor protein (AP-2) complex (Ohno et al., 1997; Boge et al., 1998), however it is not known whether this signal is functional within gp41 (Rowell et al., 1995; Boge et al., 1998).

5.5.4 Coevolution within the gp120 V3 loop

The gp120 V3 loop is critical for coreceptor binding and is also responsible in determining coreceptor usage (Hwang et al., 1991). It has also been shown to be a target for the host immune response and can somehow affect the sensitivity of virions to neutralization (Hartley et al., 2005). Recent structural analysis has shown that the V3 loop protrudes by as much as 30Å from the gp120 trimer suggesting that perhaps the N-terminus of the CCR5 receptor binds the gp120 core and V3 base whereas the V3 tip binds the coreceptor's second extracellular loop (Huang et al., 2005). We have shown the presence of coevolution both within and between amino acid residues within the V3 loop base and tip regions (Figure 5.3). There is also a large degree of coevolution involving residues within the V3 loop stem (Figure 5.3), the majority of which correlate by hydrophobicity or by both hydrophobicity and molecular weight, probably as result of the functional and structural constraints imposed on this functional domain. Resch et al. (2001) proposed that coreceptor usage is directed by positions 11 and 25 within the V3 loop (R306 and K322), and that positively charged amino acids at these positions direct CXCR4 usage while others direct CCR5 usage. These positions exhibit a high degree of coevolution, with residues both within the V3 loop (Figure 5.3) and elsewhere in *env*. For example, R306 coevolves with 21 residues within *env*, three within the V3 loop as well as residues located in the CD4 binding pocket (S364 and K432), residues the mutation of which greatly reduces CCR5 binding (Q442), residues located in the proposed minor ectodomain (E731 and G732) and residues adjacent to the second tyrosine-dependent sorting signal in gp41 (V778 and I781). K322 coevolves with eight residues in *env*, five within the V3 loop (Figure 5.3) and of the remaining three, one is located in the gp120 V2 loop (R166) and two are located within the loop region in gp41 which connects the NHR and CHR and has been associated with gp120 association (L602 and Q621).

The functions of the V3 domain are closely linked with domains elsewhere within the gp120 core (Rizzuto et al., 1998; Poignard et al., 2001; Hartley et al., 2005; Huang et al., 2005). The observation of coevolution between residues within the CD4 binding domain, residues involved in glycosylation and amino acid residues outside of the V3 loop suggested to be involved in coreceptor binding corroborates the level of evolutionary functional dependency operating within

gp120.

While previous studies have examined coevolution within the *env* V3 loop (Bickel et al., 1996; Korber et al., 1993; Gilbert et al., 2005) it is not possible to perform a direct comparison between them and this study for a number of reasons. The methods used vary between each of the studies and we have previously shown extreme differences in the sensitivities of a number of methods used to identify the presence of coevolution based on the properties of the data set (Fares and Travers, 2006).

5.5.5 Coevolution networks demonstrate the complexity of evolution operating within *env*

The extent of coevolution identified between amino acid residues throughout the *env* gene reflects the functional co-dependence of the gp120-gp41 trimer. We have shown that residues involved in both CD4 binding and in glycosylation showed a large degree of coevolution with residues throughout *env* (Figures 5.4 & 5.5). We have included amino acid residues that are directly proximal (<8Å) to functional residues in these networks as changes within proximal residues can affect the structure, and therefore functionality, of essential residues (Gloor et al., 2005). The CD4 coevolution network (Figure 5.4) contains eight residues that bind directly to CD4 as well as one residue that maps to the BMS-806 epitope, which interferes with gp120-CD4 binding (Pantophlet and Burton, 2006). All the other residues within the network are directly proximal to CD4 binding sites, five of which are residues that are directly glycosylated. Similarly, within the glycosylation network (Figure 5.5) 11 residues are directly glycosylated and one residue corresponds to part of the 2G12 epitope (Trkola et al., 1996) with the remaining residues being directly proximal to directly glycosylated residues. Of these residues proximal to directly glycosylated sites one of them when mutated interferes with chemokine receptor binding and five directly bind CD4 (Kwong et al., 1998; Rizzuto et al., 1998). In both of the networks 56% of residues are not directly functional yet exhibit a high level of coevolution within the network (Figures 5.4 & 5.5). This "backbone" of coevolving residues may maintain protein functionality through facilitating movement and communication throughout the gp120-gp41 trimer. The significant overlap of coevolving pairs between the CD4 and glycosylation networks further indicates the dependencies of evolution operating throughout the structure. Examining residues outside of the CD4 network that coevolve with direct CD4 binding residues also shows a large degree of overlap with six, four and five residues coevolving with two, three and four CD4 binding residues respectively (See supplementary information available at `http://mbe.oxfordjournals.org/cgi/content/full/msm213/DC1`). Our results are overlapping with those presented by Poon and coworkers, where they detect coevolution between N-linked glycosylated sites in the envelope protein of HIV-1 using a phylogenetic and a Bayesian

graphical models of evolution (Poon et al., 2007). Proximal coevolving amino acid sites can also indicate compensatory epistatic effects. Compensatory mutations have been observed in the HIV genome with the majority associated in *pol* with drug resistance (Piana et al., 2002; Menendez-Arias et al., 2003; Perno et al., 2006) as well as a number of compensatory mutations identified in *gag* (Friedrich et al., 2004; Yeh et al., 2006). A recent study by Gorry and colleagues, however, proposed the presence of compensatory mutations between residues 308R/317F and 308R/321G that affect coreceptor binding (Gorry et al., 2007). Our study did not observe direct covariation/coevolution between either of these pairs although residues proximal to both of these pairs were observed as coevolving (Figure 5.3). Similarly, Baldwin and Berkhout proposed a number of potential compensatory mutations in both gp120 and gp41 which enabled escape from T20-dependent replication (Baldwin and Berkhout, 2006). The initial mutations that caused T20 dependency occurred as V549A and N637K. Multiple occurrences of a G431R mutation enabled T20-dependency escape were observed suggesting that compensatory mutations within the CD4 binding domain may affect T20 dependency. We have observed high levels of coevolution within the CD4 binding domain (Figure 5.4) and have also observed strong coevolution between 430V located in the CD4 binding domain and 567Q identified by Baldwin and Berkhout as an escape mutant from T20-dependent replication (Baldwin and Berkhout, 2006). In addition to proximal coevolving glycosylated sites, we observed many of the coevolving pairs of sites to present distances in the structure above 4.5 Å. Coevolution between distant N-glycosylated sites may be convenient to ensure an efficient shielding through glycosylation of sites recognized by the host defense system as previously pointed out (Poon et al., 2007).

Although detection of coevolution is an interesting problem *per se*, the pragmatic value of detecting coevolution transcends many areas of research. The understanding of the molecular communication between the different proteins involved in infectivity and spread in HIV-1 is essential to identify functionally/structurally important protein domains and hence to design proper therapeutics against the virus. This communication is only tractable from the evolutionary point of view and in this sense coevolution analysis can easily highlight such dependencies. In this study we aimed at identifying these covariation dependencies in order to understand the evolutionary dynamic of the two most important proteins of the HIV-I infection machinery.

While we have been able to assign biological significance to many of the coevolving pairs and groups identified in this study, this is not always the case. Many of the residues in the gp120-gp41 trimer, while not directly functionally important, may be important in the maintenance of the protein in a functional conformation or may be involved in intra- or inter-protein communication. Highly significant coevolving residues may provide ideal targets for future site-directed mutagenesis analysis in the identification of functional domains and have also been suggested as a strategy for the design of broadly neutralizing vaccines (Korber et al., 1993).

We propose that pairs/groups of coevolving amino acids are seen across the entire HIV-1

group M phylogeny but also that subtype-specific pairs/groups exist. In fact subtype specific patterns of evolution have been previously identified (Korber et al., 1994; Gaschen et al., 2002). The association of these subtypes specific evolutionary patterns and the structure characteristics of the protein have been elegantly examined in a recent work (Gnanakaran et al., 2007). However, coevolutionary patterns in specific subtypes have to be as yet comprehensively studied. Residues observed as coevolving across group M have been functionally/structurally constrained throughout the evolution of group M while subtype-specific coevolving residues may represent novel dependencies within *env* for a particular subtype. Analysis of such subtype specific dependencies may provide clues as to subtype-specific mechanisms immune escape or infectivity.

5.6 Acknowledgements

We thank Peter Kwong for supplying the coordinates of the gp120 trimer and Christopher J. Creevey, Jennifer Commins and Christina Toft for assistance with the software used to produce Figure 5.2. This study was supported by Science Foundation Ireland.

Chapter 6

Understanding the genetic diversity of HCV

6.1 Abstract

Hepatitis C virus is a rapidly evolving RNA virus that exhibits enormous genetic diversity and is a major cause of liver disease worldwide. It is classified into seven genotypes and numerous subtypes which differ in diversity, transmission, geographical distribution and in response to treatment. Despite a considerable amount of progress made on understanding the disease there is no vaccine and current antiviral therapy is relatively toxic. Contrary to this, studies amalgamating new molecular sequence data remain unperformed. Here, we explore for the first time using a dataset comprising 133 complete genomes the complex evolutionary pressures that underlie the evolution of HCV genotypes. We show that the frequency of recombination is rare and plays a very minor role in the evolution of HCV variants. Similarly, screening of isolates revealed molecular signatures for positive Darwinian selection in ancestral lineages leading to genotypes. The fixation of such adaptive mutations was predominantly found in the structural genes that encode the capsid and envelope glycoproteins while non-structural proteins are enriched by purifying selection. Furthermore analyses of synonymous sites and slightly deleterious mutations incriminate genetic drift in accounting for the observed sequence diversity between types. However taken together these patterns support that varied epidemiological behaviour of HCV types and subtypes is largely defined by transmission route rather than any common evolutionary mechanism acting on strains. These results support that given an efficient route of transmission any of the known HCV types could 'emerge' and generate a future epidemic.

6.2 Introduction

Hepatitis C virus (HCV) is a single stranded RNA virus of approximately 9400 kb and is the leading etiologic agent of chronic liver disease. The positive-sense genome of HCV contains a single large ORF that encodes a polyprotein containing 10 mature proteins. The structural proteins, the core (C) and envelope glycoproteins E1 and E2 are in the N-terminal part of the polyprotein and form the virion. The nonstructural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B) have various functions and form the replication complex (Roingeard et al., 2004). According to WHO, the global prevalence of HCV infection is estimated to be 3%, representing about 180 million people (WHO, 1999)with 70-85% of infected adults developing chronic infections (Alberti et al., 1999). HCV is classified in the family *Flaviviridae (Choo et al., 1989),* a family of positive sense RNA viruses that evolve rapidly due to high rates of mutation and replication and lack of proofreading mechanisms (Jenkins et al., 2002) although it differs from the original vector-borne members of the family.

Phylogenetic analyses of HCV genomes reveals that sequences fall into different clusters. This observation resulted in a nomenclature that recognises different hierarchies of divergence (Simmonds et al., 2005). Six distinct virus types have been described that differ from each other by more than 30% over the complete virus genome (Simmonds et al., 1993; Simmonds, 2004). These genotypes are further divided into multiple epidemiologically distinct subtypes (labelled a, b, c etc, in order of discovery) differing by 20-25% from one another, while within each subtype variation is less than 10% . Despite substantial sequence variation all genotypes share the same linear structure of genes of nearly identical size.

The predominant genotype in most areas and probably the most studied is genotype 1. Thirteen subtypes of HCV genotype 1 have been described so far (from 1a to 1m). However, only three (1a, 1b and 1c), for which the complete genome sequence has been obtained, have the status of confirmed subtype. The remaining subtypes (from 1d to 1m), from which only partial sequences are known, have been denoted as provisional. However, very recently this has been amended with the description of a complete genotype sequence of a HCV subtype 1g (Bracho et al., 2008). In addition, a complete genotype 1 sequence from an Equatorial Guinea isolate with unassigned subtype is also available (Bracho et al., 2006a). Some, such as genotype 1a and 1b, have become widely distributed as a result of transmission through blood transfusion and needle sharing between infected drug users (IDUs) (Simmonds, 2004). Genotype 2 is predominantly found in western Africa with previous studies reporting great genetic diversity associated with its high virus prevalence, suggesting that this genotype has been present in the human populations in west African countries for some time. Previous studies indicate the striking different epidemic histories of different subtypes. Recently, a comparative genetic analysis of HCV genotype 2 infection in West Africa and Cameroon showed that the virus was spreading

relatively slowly within the West African population from 1630 to 1900, whilst in Cameron there was rapid transmission from 1920 to 1960 (Pouillot et al., 2008; Njouom et al., 2006). Genotype 3 is endemic in some parts of Asia such as the Northern Indian subcontinent but subtype 3a similarly to 1a and 1b has become distributed very widely as a results of IDUs particularly from Europe. Genotype 4 is found prevalent in the Middle East, including Egypt, central Africa and Cameroon. However, it has also been found at increasing frequencies among southern Europeans mainly among IDUs. Despite the genotype being divided up into 18 subtypes (a-u) only 5 subtypes have complete genomes described. The predominance of the HCV epidemic in Egypt has been linked to extensive parenteral treatment of schistosomiasis in the 1950s where the rapid exponential growth of infections coincide with the extensive injection campaigns (Pybus et al., 2003a; Ray et al., 2000). Genotype 5 was thought to be an uncommon HCV genotype and believed to be confined to the northern part of South Africa and was considered as anecdotal elsewhere. However, the epidemiology of this HCV genotype seems to be more diverse than first thought. There have been sporadic outbreaks of the disease in United Kingdom, the Netherlands, Ireland, Australia, Canada, Brazil, and Germany (Davidson et al., 1995; Levi et al., 2002; Ross et al., 2000; Murphy et al., 1994). More recently it was found at an unusually high prevalence in an area in Belgium where strains were found to be circulating for quite a long time (approx 120 years) (Verbeeck et al., 2006). Although it is still not clear where the region of endemic infection for genotype 5 is located (Verbeeck et al., 2008). Geographically genotype 6 variants have been identified exclusively in south-eastern Asia or immigrants from this region. This genotype shows the greatest genetic diversity with viruses classified into 19 subtypes (6a-6q, 6t, 6u) based on complete genomic analysis. Two provisionally assigned subtypes (6r, 6s) were proposed based on sequence analysis of core/E1 and NS5B from strains isolated from Cambodian immigrants in Canada (Murphy et al., 2007b). In addition a number of variants isolated from China cannot be classified and remain unassigned (Lu et al., 2006b). Almost all subtypes of genotype 6 (except for subtype 6a which has a high prevalence) appear to be relatively rare samples from a restricted geographical region and highly genetically divergent from each other. This suggests a long term endemic transmission of variants in south-eastern Asia with little exchange of virus lineages with outside populations. Perhaps such a scenario is consistent with a zoonotic origin of HCV genotype 6 as previously described for genotypes belonging to hepatitis E virus (Lu et al., 2006a). Nevertheless, what roles played during virus evolution and transmission to create such great genetic diversity remains unknown.

More recently a group of researchers announced at the 58th Annual Meeting of the American Association for the Study of Liver Diseases (AASLD 2007) in Boston, USA the apparent discovery of a new seventh HCV genotype originating in Central Africa (Murphy et al., 2007a). While unpublished at this time the investigators identified 3 HCV variants (QC69, CS101285, and CS101300) that clustered together, but did not fall within the classification of the 6 known

genotypes. QC69 was isolated from a patient residing in Canada, while CS101285 and CS101300 were obtained from patients in Belgium. However, all 3 patients came from the Democratic Republic of Congo, where they were presumably infected. With two partial coding region sequences and a complete genome determined, phylogenetic analysis have shown that this genotype has a close relationship with genotype 2 variants, analogous to the relationship between genotype 1 and 4 variants and is not a product of recombination between known genotypes. This single genome was also included in this study merely for phylogenetic purposes as it does not contain sufficient information to carry out selection analysis on. Nevertheless, the finding of this new genotype in Central Africa may contribute in elucidating the origin of the worldwide HCV epidemic.

Due to the paucity of sequence data available for some genotypes a huge sequencing initiative supported by the NIH to improve the HCV sequence database has been promoted. This has resulted in the release of an abundance amount of complete genomic sequence, most of which belong to genotype 6 subtypes (Lu et al., 2007a, 2008, 2006b, 2007b; Noppornpanth et al., 2008) and genotype 4 (Timm et al., 2007; Hmaied et al., 2007; Kuntzen et al., 2008). In addition, further subtypes have been characterised for genotype 1 (Bracho et al., 2008) and as yet an uncharacterised distinct genotype 1 variant isolated from Equatorial Guinea (Bracho et al., 2006a). Increasing amounts of molecular sequence data enables more comparisons among viruses and between viruses and other organisms. Thus, this provides copious grist for understanding the constraints that shape HCV evolution and may bring forth new strategies for public health control and clinical treatment of HCV infections. To date there has been no single study focusing on the evolution of all HCV genotypes. In fact the evolutionary pressures of HCV and the diversification of genotypes still remains much a mystery. Moreover it is broadly conceived that neutral genetic drift is the major driving force of HCV evolution over time and this process of accumulating sequence changes neutrally accounts for much of the variability observed between the geographically isolated genotypes of HCV. Importantly, given the growing evidence for genotype-specific differences in persistence and in response to antiviral treatment no evolutionary study has been performed to investigate if there is such a confounding factor that varies in a subtype-specific manner. Similarly, the role and extent that sequence drift plays has never been fully evaluated across all six genotypes. Likewise quantifying the role that slightly deleterious mutations and other constraints such as recombination and coevolution play in genotype evolution is fundamental to understanding HCV heterogeneity. With all this in mind, we assemble the most comprehensive dataset of HCV complete genomes and explore the main evolutionary forces shaping HCV dynamics across six genotypes.

6.3 Materials and Methods

6.3.1 Genomic Data

Complete genome sequences representing all six genotypes of HCV were downloaded from GenBank (See Table 1 in electronic appendix for a description). In total a 133 full length genome based alignment comprising 52, 12, 6 19, 2, 41 and 1 sequences for genotype 1, 2, 3, 4, 5, 6 and 7 respectively was used. We aligned protein sequences using the program Muscle v3.6 (Edgar, 2004) with the default settings. We then aligned nucleotide sequences concatenating triplets of nucleotides according to the protein sequence alignment. The total length of the alignment was 9213 nucleotides.

6.3.2 Detection of selective pressures at single amino acid site and individual lineages

To determine the gene and site specific selection pressures acting on HCV genotypes, we estimated the rate of of non synonymous (dN) to synonymous (dS) substitutions per nucleotide site. The ratio between the two rates $\omega = d_N/d_S$ helps us to elucidate if a gene has been fixing amino acid replacements neutrally ($\omega = 1$), under adaptive evolution ($\omega > 1$) or purifying selection ($\omega < 1$) as described in section 1.5.2. Although it has been shown that ω is a poor indicator of adaptive evolution (Sharp, 1997; Crandall et al., 1999) as the detection of episodic position selection may be swamped by strong purifying selection against the majority of amino acid mutations throughout most of the evolutionary time resulting in a $\omega << 1$. However, one way to overcome this problem is to estimate ω for specific branches in a given phylogeny and for each region of a protein. Nevertheless, in general ω remains a conservative detector of adaptive evolution.

Here, we focus on positively selected sites that were inferred by using several codon substitution models implemented in the program Codeml in the PAML package v3.15 (Yang, 1997) and we determine the extent of overlap in terms of selection intensity and location among different HCV lineages. We have carried out this analysis by comparing each individual genotype to all other genotypes in an attempt to identify amino acids whose evolutionary history appears to be unique in terms of selective constraints for that genotype. The identification of such sites yields information as to unique subtype-specific molecular traits that may also manifest as unique biological traits. Therefore in order to identify different pathways of evolution taken by genes we analysed those evolutionary forces acting on each individual gene for each genotype. We first quantified positive selection in all 10 HCV genes for the six genotypes (1, 2, 3, 4, 5, 6) and tested whether phylogenetically divergent lineages are subjected to similar selective pressures. We used a model that assumes a single ω value for all lineages and sites against the occurrence

of different categories of ω values per site shared across all lineages. This is implemented in the program Codeml from the PAML package v3.15 (Yang, 1997) as M0 and M3 respectively. The outcome of M3 was then compared with a branch site-specific models (model B), which allows for rate heterogeneity across sites and across the tree. Model B is an extension of M3 as long as just two categories are considered (Yang et al., 2000) and allows us to have some sites under a particular ω value potentially different to the ω considered for the rest of the tree (Zhang et al., 2005; Yang and Nielsen, 2002). Since the branch site models operate by allowing the user to examine the evolutionary mechanisms occurring in a particular lineage in the tree (the foreground) against the other lineages (the background), they provide a unique mode of analysis by allowing the selective constraints operating on certain sequences in a genotype to be compared to the selective constraints operating on all the other sequences present in the data set. The branches leading to genotypes 1-6 were labelled and to ensure accurate results, each model was run three times using different starting ω values. Also, any putative recombinants detected were excluded from this part of the analysis. The significance of positive selection can be confirmed with a likelihood ratio test (LRT) between null models and those able to account for positive selection. An LRT is performed by taking twice the difference of log likelihood between nested models and comparing the result to a χ^2 distribution with degrees of freedom equivalent to the difference in the number of parameters between the models. Models compared in this study using LRT were M0 and M3 and M3 and model B (for more information on the models used see Yang and Nielsen (2002); Zhang et al. (2005)).

6.3.3 Recombination analysis

In order to test for the presence and extent of recombination we used two different strategies. One approach to assessing the extent of recombination in HCV was to document cases in which different genes produced phylogenetic trees with different topologies – so-called phylogenetic incongruence. In this case we inferred maximum likelihood phylogenetic trees from PAUP* (Swofford, 1998) using the best fit model of nucleotide substitution for each gene determined using ModelTest (Posada and Crandall, 1998). Although this test is conservative our primary aim is to identify individual sequences that changed phylogenetic position, thus demonstrating that genes show discordant inter-genotype relationships as intra-typic relationships are not completely described in the literature. To explore the extent of recombination further and to identify putative recombinant genomes we used the RDP3 package version 3.22β (`http://darwin.uvigo.es/rdp/rdp.html`). Once all putative recombination events are identified the results are sorted and the number of unique recombination events identified with the daughter, parental sequences and breakpoints determined. This package incorporates an exhaustive list of published recombination detection methods into a single suite of tools (Mar-

tin et al., 2005c). The programs used in this study were: RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), BOOTSCAN (Martin et al., 2005a), MAXCHI (Smith, 1992), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs et al., 2000). In all cases, default parameters were used. In order to ensure our approach remains conservative we considered only events predicted by half of the methods as significant.

6.3.4 Identifying Constraints at Synonymous Sites

It has been shown that stability of RNA molecule secondary structure as well as translational selection may impose constraints on synonymous sites leading to lower d_S values and consequently to inflated ω estimates (Resch et al., 2007; Parmley et al., 2006; Mayrose et al., 2007; Chamary et al., 2006). To determine the degree of relaxed selective constraints in each one of the proteins for each HCV genotype, we estimated d_S by the method of Li (1993) as implemented in SWAPSC. This is a sliding window method that dissects the selective constraints operating on different regions in a protein-coding gene (Fares, 2004). Due to the limited evolutionary signal contained in one codon site, we slid a window of 20 codons along the genome and estimated d_S for each sliding step. We then tested whether d_S was homogeneously distributed along the genome, as it should be expected under a neutral evolution model for synonymous sites. In particular, we were interested on knowing if d_S accumulates differently in proteins and whether this effect is genotype-specific.

6.3.5 Identifying and Quantifying Slightly Deleterious Mutations

To quantify the number of SDMs we developed a new and simple method. Briefly, we first identified sites under strong purifying selection (highly conserved sites in the alignment). We then calculated the amino acid transition scores for the pair wise sequence comparisons at these sites. We calculated transition scores following the appropriate BLOck SUbstitution Matrices (BLOSUM) (Henikoff and Henikoff, 1992) given the average pairwise distance for the alignment. Once the distribution of BLOSUM values was determined for that site in the alignment, we tested each genotype for the accumulation of SDMs by identifying lineages accumulating changes with transition values showing significant values (highly negative transition scores) when compared to the distribution of values of the rest of the alignment. The rationale behind this analysis is that highly conserved amino acid sites (for example, those amino acid sites showing BLOSUM values highly positive or close to 0) are expected to be functionally important because of their high level of conservation. A strong change at these sites in particular lineages (for example at terminal lineages within a genotype) is more likely to be a SDM fixed by genetic drift than an adaptive change. Following this procedure, we identified SDMs in each one of the proteins at lineages leading to each one of the genotypes and within genotypes. As in all the cases, we

normalised the proportion of sites showing evidence of being SDMs as to make possible the direct comparison of these proportions between genomic regions and genotypes. This normalisation consisted of the transformation of these per-gene and per-serotype proportions in the proportion taking into account genotypes and genes together.

6.3.6 Codon Usage Analysis

The codon adaptation index (CAI) was proposed as a quantitative way of predicting the expression level of a gene based on its codon sequence (Sharp et al., 1993). By using highly expressed host genes as the reference set we can predict the level of expression of viral genes in human cells (Haas et al., 1996). This highly expressed codon set has been used successfully for codon optimisation in viral genes (Kadam and Ghosh, 2008; Sewatanon et al., 2007). Hence, the closer the value to 1.00, the more the gene is adapted for effective translation in human cells (Sharp and Li, 1987). The CAI was calculated on the server created and maintained by the Freeland Research Group at the Department of Biological Sciences, University of Maryland, USA (`http://www.evolvingcode.net/codon/cai/cai.php`). This test was performed in order to evaluate whether the different gene evolutionary dynamics were as a result of varying expression levels.

6.4 Results

6.4.1 Heterogeneous selective constraints

Maximum likelihood estimates of parameters under different models of variable ω among sites and branches are presented in table 6.1. We calculated the likelihood of the data under M0, M3 and the branch site codon model B. The branch site models were implemented to detect any sites that have evolved uniquely to a particular genotype when compared to the other genotypes. In this instance the foregrounds branches were specified *a priori* as described in materials and methods to test whether adaptive evolution is driving genotype diversity after its divergences. In all but a few of the comparisons across genes, M3 significantly improved the log-likelihood value with respect to M0 (Table 6.1). This indicates that HCV variants have fixed changes under heterogeneous evolutionary constraints across codons. To test for the presence of heterogeneous constraints and adaptive evolution in HCV lineages, we compared the branch site-specific model B for each one of the lineages with the M3 model. In almost all instances, Model B significantly improved the log-likelihood value of the M3 model. Throughout all proteins there was evidence for strong purifying selection in the background lineages (ω varying from 0.003 to 0.03). For those lineages that we were interested in i.e. the foreground branches, purifying selection was dominating. However, some genotypes showed varying degrees of selective constraints in genes.

The nucleocapsid protein for two genotypes, 1 and 3 showed signs of positive selection with $\omega > 1$ (Figure 6.1) while genotypes 2 and 6 appear to be evolving neutrally ($\omega = 1$). Similarly, the envelope that encodes two proteins, E1 and E2, which are possibly responsible for the binding and entry of the virus to target cells, showed evidence for positive selection in genotypes 3, 4, 5 and 6 (Figure 6.2 & 6.3). Positive diversifying selection appeared more sporadic in non-structural proteins with the hydrophobic NS4B protein under positive selection for genotype 3 (Figure 6.4). The remaining proteins in the genome did not reveal any strong evidence for positive selection but strong purifying selection. However, the first non-structural protein encoded, NS2 revealed that genotype 3 is evolving towards neutrality ($\omega = 1$). Also it must be stated that any signs of positive selection found in the p7 and NS4A regions were discounted due to the small size and lack of phylogenetic signal in both proteins. Results for NS5B are not shown as they have never finished as a result of the severe computational requirement. However, while there is statistical evidence for the presence of sites under positive selection on a lineage there is not sufficient information to allow the Bayes empirical Bayes (BEB) method to identify sites. This problem is not just confined to this analysis and has been discussed elsewhere (Zhang et al., 2005)

Table 6.1: Likelihood ratio test for the comparsion between codon based models. M0 was compared to M3. Also, the M3 model was compared to branch-site model for each branch leading to a major HCV clade using the likelihood ratio test. Twice the difference in the log-likelihood values for nested models were approximated to a chi-squared distribution with the degrees of freedom the difference in the number of parameters between the models compared. NS refers to those tests that were not significant

Models	***2ΔlnLk***	***Δ df***	***p-value***
Core			
M0 vs M3	618.44	4	0
Model B (Genotype 1) vs M3	18.59	2	$9.20 \cdot 10^{-5}$
Model B (Genotype 2) vs M3	22.20	2	$1.50 \cdot 10^{-5}$
Model B (Genotype 3) vs M3	39.71	2	0
Model B (Genotype 4) vs M3	23.06	2	$1.00 \cdot 10^{-5}$
Model B (Genotype 5) vs M3	18.78	2	$8.4 \cdot 10^{-5}$
Model B (Genotype 6) vs M3	8.922	2	$1.1 \cdot 10^{-2}$
E1			
M0 vs M3	1470.03	4	0
Model B (Genotype 1) vs M3	86.71	2	0
Model B (Genotype 2) vs M3	109.21	2	0
Model B (Genotype 3) vs M3	21.48	2	$2.20 \cdot 10^{-5}$
Model B (Genotype 4) vs M3	52.60	2	0

Models	***2ΔlnLk***	***Δ df***	***p-value***
Model B (Genotype 5) vs M3	24.09	2	$6.00 \cdot 10^{-6}$
Model B (Genotype 6) vs M3	115.60	2	0
E2			
M0 vs M3	6592.82	4	0
Model B (Genotype 1) vs M3	712.09	2	0
Model B (Genotype 2) vs M3	121.47	2	0
Model B (Genotype 3) vs M3	91.94	2	0
Model B (Genotype 4) vs M3	133.51	2	0
Model B (Genotype 5) vs M3	24.91	2	$4.00 \cdot 10^{-6}$
Model B (Genotype 6) vs M3	245.19	2	0
p7			
M0 vs M3	341.13	4	0
Model B (Genotype 1) vs M3	22.45	2	$1.30 \cdot 10^{-5}$
Model B (Genotype 2) vs M3	41.19	2	0
Model B (Genotype 3) vs M3	NS	2	-
Model B (Genotype 4) vs M3	20.60	2	$3.40 \cdot 10^{-5}$
Model B (Genotype 5) vs M3	15.94	2	$3.46 \cdot 10^{-4}$
Model B (Genotype 6) vs M3	27.78	2	$1.00 \cdot 10^{-6}$
NS2			
M0 vs M3	1860.95	4	0
Model B (Genotype 1) vs M3	181.4	2	0
Model B (Genotype 2) vs M3	49.91	2	0
Model B (Genotype 3) vs M3	42.26	2	0
Model B (Genotype 4) vs M3	169.0	2	0
Model B (Genotype 5) vs M3	NS	2	-
Model B (Genotype 6) vs M3	296.39	2	0
NS3			
M0 vs M3	287.54	4	0
Model B (Genotype 1) vs M3	305.70	2	0
Model B (Genotype 2) vs M3	81.91	2	0
Model B (Genotype 3) vs M3	54.45	2	0
Model B (Genotype 4) vs M3	62.61	2	0
Model B (Genotype 5) vs M3	13.59	2	$1.10 \cdot 10^{-3}$

Models	*2ΔlnLk*	*Δ df*	*p-value*
Model B (Genotype 6) vs M3	219.38	2	0
NS4A			
M0 vs M3	296.28	4	0
Model B (Genotype 1) vs M3	44.04	2	0
Model B (Genotype 2) vs M3	20.49	2	$3.60 \cdot 10^{-5}$
Model B (Genotype 3) vs M3	9.43	2	$8.96 \cdot 10^{-3}$
Model B (Genotype 4) vs M3	9.05	2	$1.08 \cdot 10^{-2}$
Model B (Genotype 5) vs M3	NS	2	-
Model B (Genotype 6) vs M3	24.05	2	$6.00 \cdot 10^{-6}$
NS4B			
M0 vs M3	1116.17	4	0
Model B (Genotype 1) vs M3	107.79	2	0
Model B (Genotype 2) vs M3	26.59	2	$2.00 \cdot 10^{-6}$
Model B (Genotype 3) vs M3	25.26	2	$3.00 \cdot 10^{-6}$
Model B (Genotype 4) vs M3	29.23	2	0
Model B (Genotype 5) vs M3	NS	2	-
Model B (Genotype 6) vs M3	74.17	2	0
NS5A			
M0 vs M3	3785.27	4	0
Model B (Genotype 1) vs M3	NS	2	-
Model B (Genotype 2) vs M3	175.92	2	0
Model B (Genotype 3) vs M3	258.32	2	0
Model B (Genotype 4) vs M3	304.11	2	0
Model B (Genotype 5) vs M3	243.29	2	0
Model B (Genotype 6) vs M3	NS	2	-

6.4.2 Recombination

Three recombination events were detected using the RDP program in genotype 1 with each event identified in Khaja 1, a variant belonging to subtype 1c (Table 6.2). No recombination was reported in any other genome sequence for the remaining genotypes. The confinement of recombination to genotype 1 represents a 1a/1c intragenotypic recombinant. The putative breakpoints detected are predominantly found in the structural regions of the genome and con-

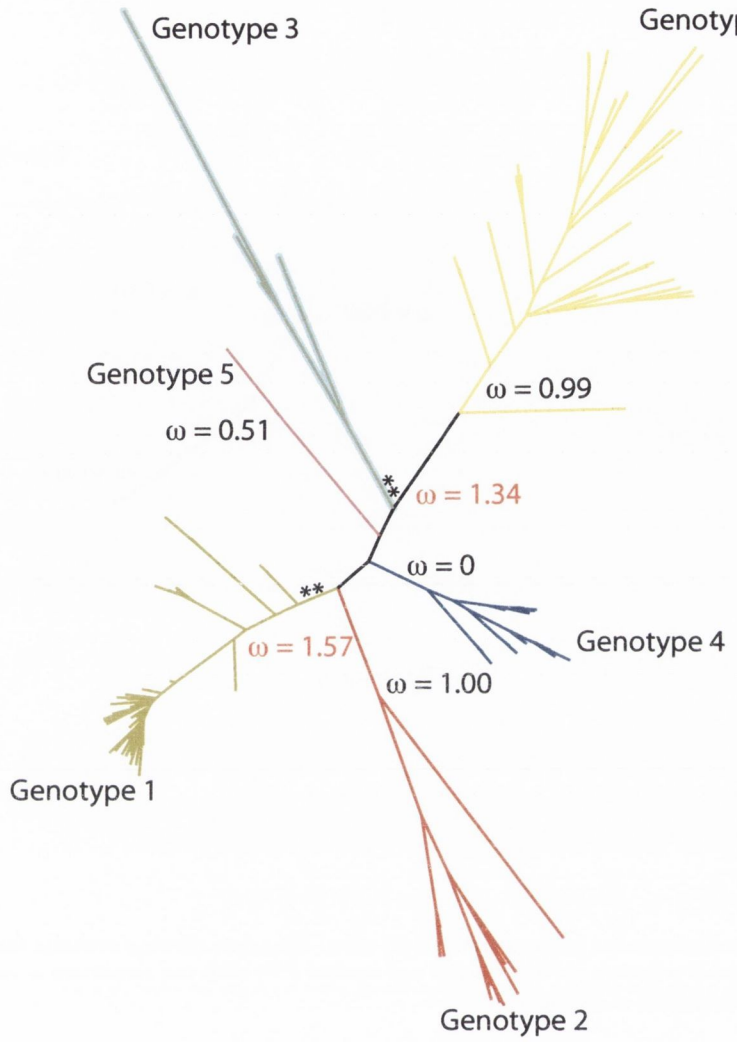


Figure 6.1: Constructed phylogeny of the core gene. Branches showing evidence for the presence of sites under positive selection on that lineage are marked (**) with the respective ω value in red. All other branches have their ω shown.

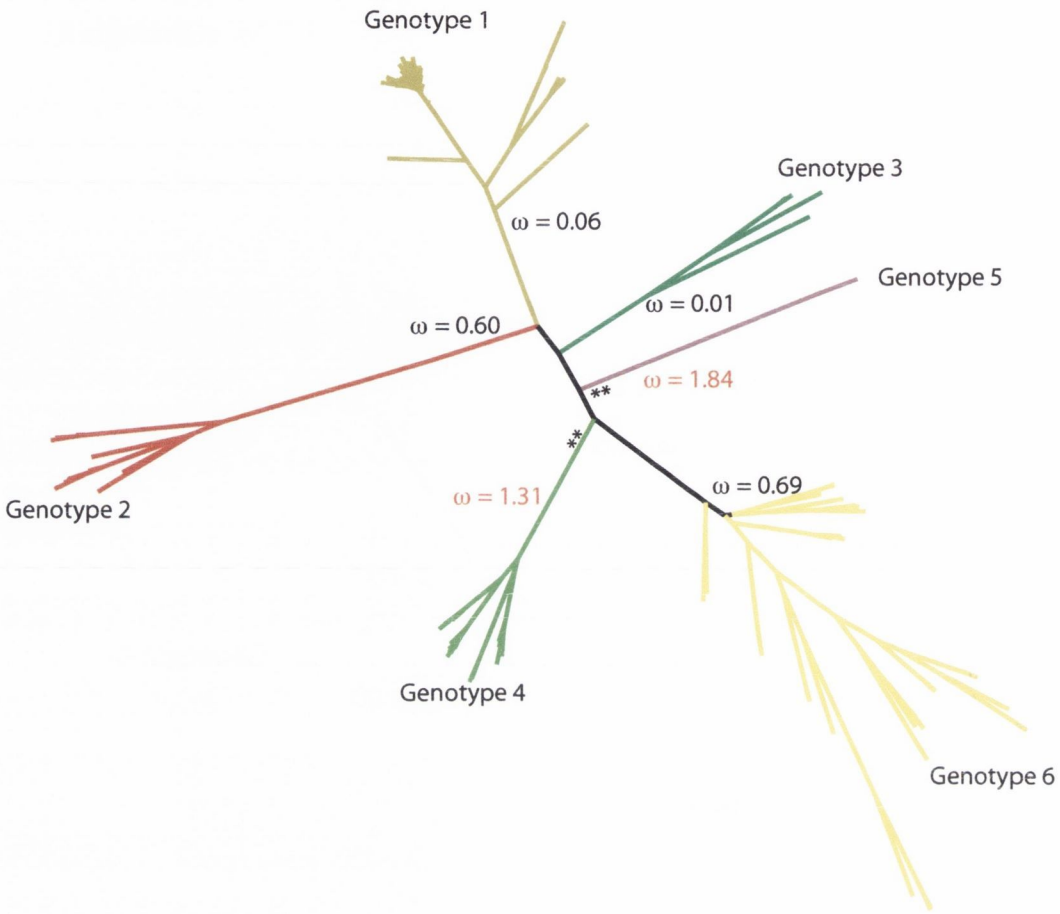


Figure 6.2: Constructed phylogeny of the E1 gene. Branches showing evidence for the presence of sites under positive selection on that lineage are marked (**) with the respective ω value in red. All other branches have their ω shown.

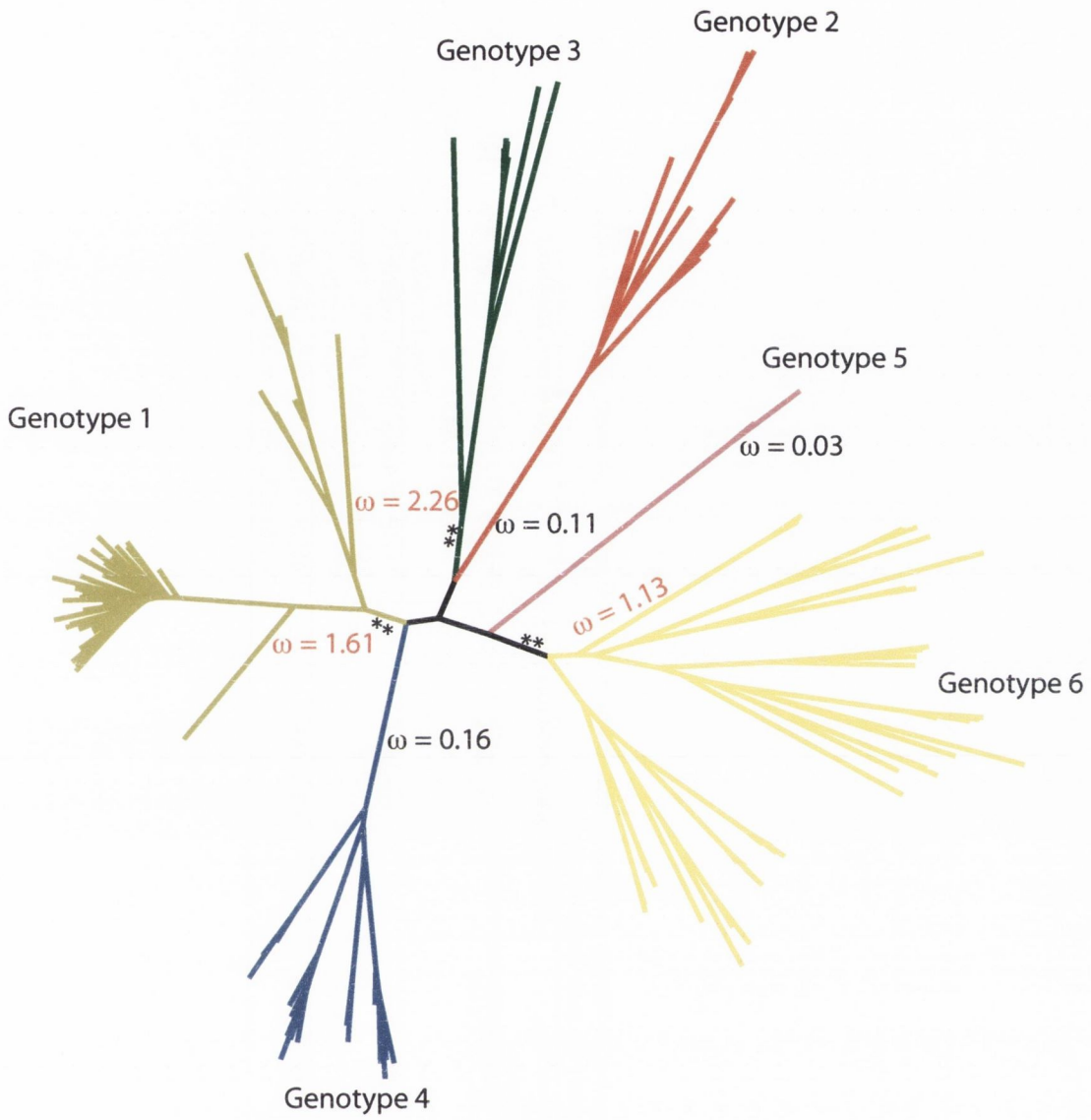


Figure 6.3: Constructed phylogeny of the E2 gene. Branches showing evidence for the presence of sites under positive selection on that lineage are marked (**) with the respective ω value in red. All other branches have their ω shown.

Table 6.2: Recombination events inferred using a number of different methods contained within the RDP software package. Only events supported by at least three different methods are reported. The support probability for each method is shown.

Recombinant virus (subtype)	**Putative parentals (major, minor)**	**Breakpoint location (nt)**	**Inference method (*P* value)**
Khaja -1 (1c)	AY051292 (1c) , HCVPT (1a)	2740-3478	RDP (1.541×10^{-65}), GENECONV (1.776×10^{-42}), MaxChi (1.747×10^{-23}), Chimaera (6.903×10^{-19}), 3Seq (6.833×10^{-23})
Khaja -1 (1c)	AY051292 (1c) , H77 (1a)	969-1874	RDP (1.538×10^{-58}), GENECONV (5.932×10^{-40}), MaxChi (1.139×10^{-16}), Chimaera (1.428×10^{-18}), 3Seq (1.719×10^{-23})
Khaja -1 (1c)	AY051292 (1c) , PHCV1SF9 (1a)	9010-569	RDP (7.794×10^{-37}), GENECONV (4.023×10^{-26}), MaxChi (3.734×10^{-9}), Chimaera (8.018×10^{-10})

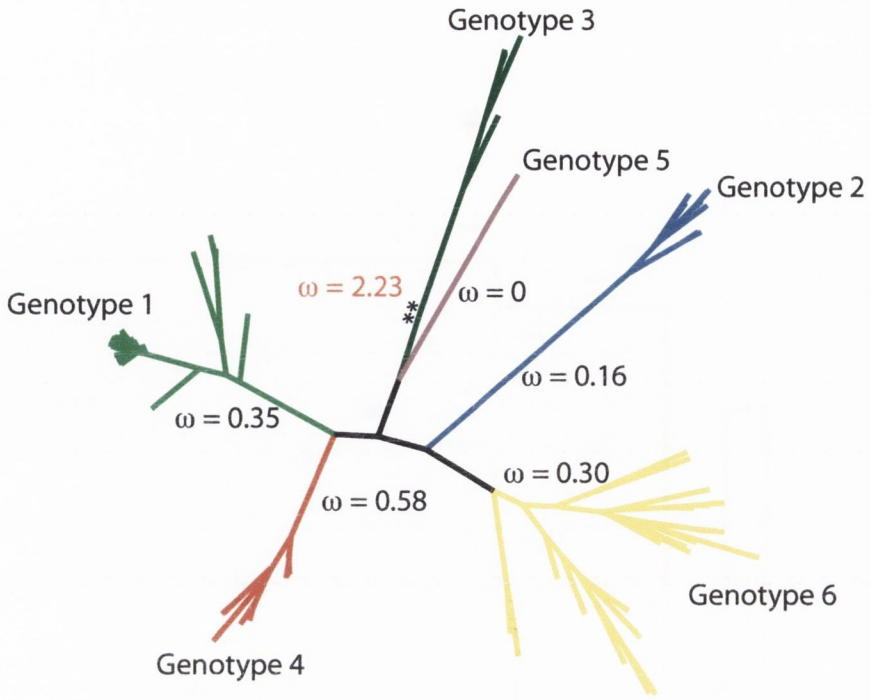


Figure 6.4: Constructed phylogeny of the NS4B gene. Branches showing evidence for the presence of sites under positive selection on that lineage are marked (**) with the respective ω value in red. All other branches have their ω shown.

firmed from conflicting phylogenetic signals. Phylogenetic trees inferred for the Core and E2 genes show that the location of Khaja-1 was incongruent, appearing with the group 1a clade (Figure 6.5 and 6.6). No discordant relationships were detected for any other genotypes in the remaining genes.

6.4.3 Homogeneity of synonymous sites

Analysis of d_S shows that in all genotypes its distribution was highly homogeneous throughout the genome with no differences observed in structural proteins versus non-structural proteins. The frequency of silent substitutions was constant over most of the polyprotein open reading frame, with one exception (See figure 6.7). Genotype 5 has a strikingly higher synonymous substitution rate in its structural proteins than the rest of the genome and a lower synonymous substitution rate in its first non-structural proteins encoded. This difference is probably due to more relaxed constraints on synonymous sites of structural proteins compared to non-structural proteins. In fact the overall pattern that we observe in this genotype is different to that in other genotypes but due to the lack of complete genomes available (only two), this may represent an artifact.

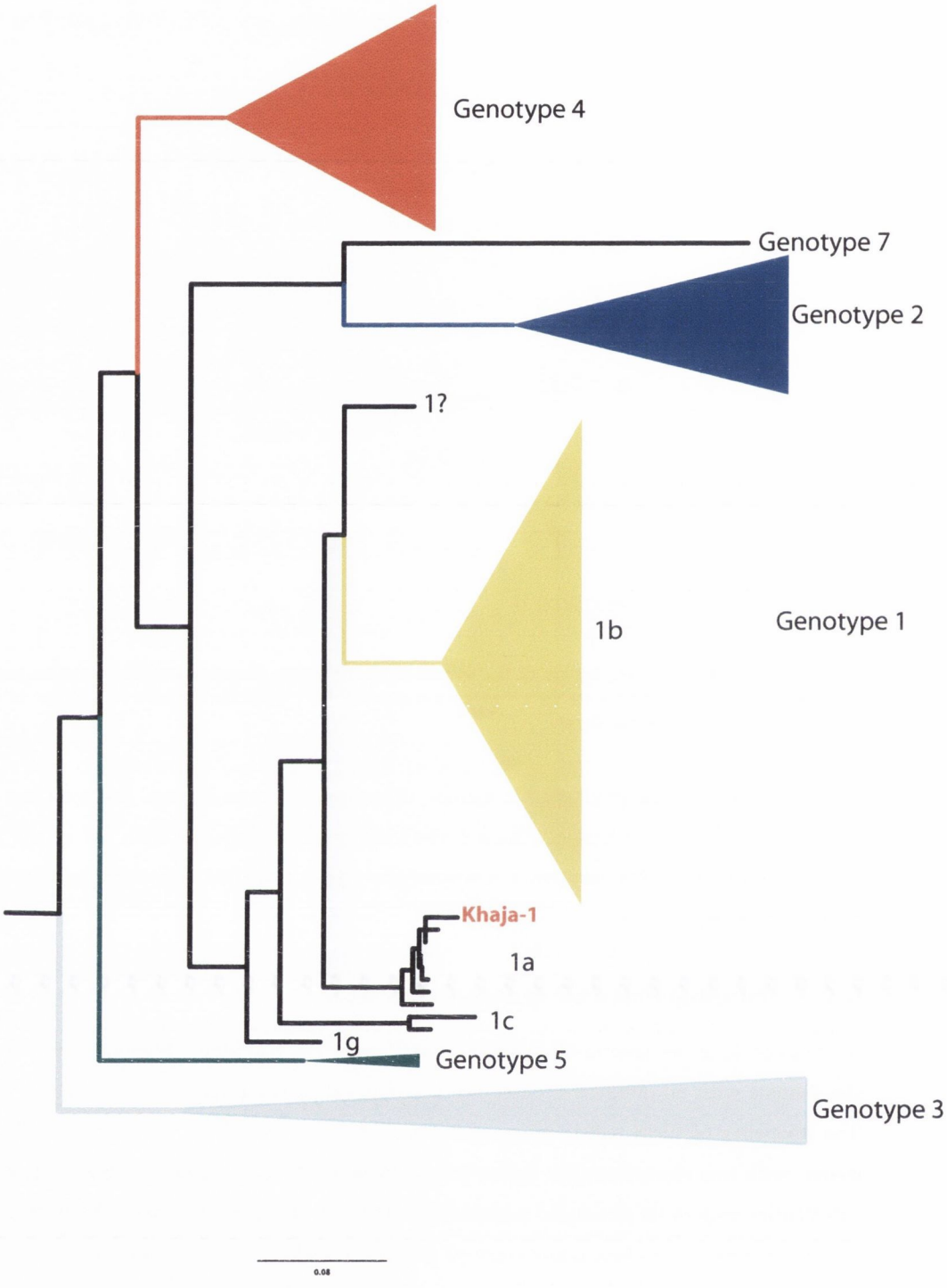


Figure 6.5: Maximum likelihood phylogenetic tree of the core gene depicting the topology of HCV isolates. For clarity purposes genotype 6 is not shown as the main focus is on genotype 1 with its putative recombinant highlighted in red. Each clade is labelled accordingly with respect to its genotype.

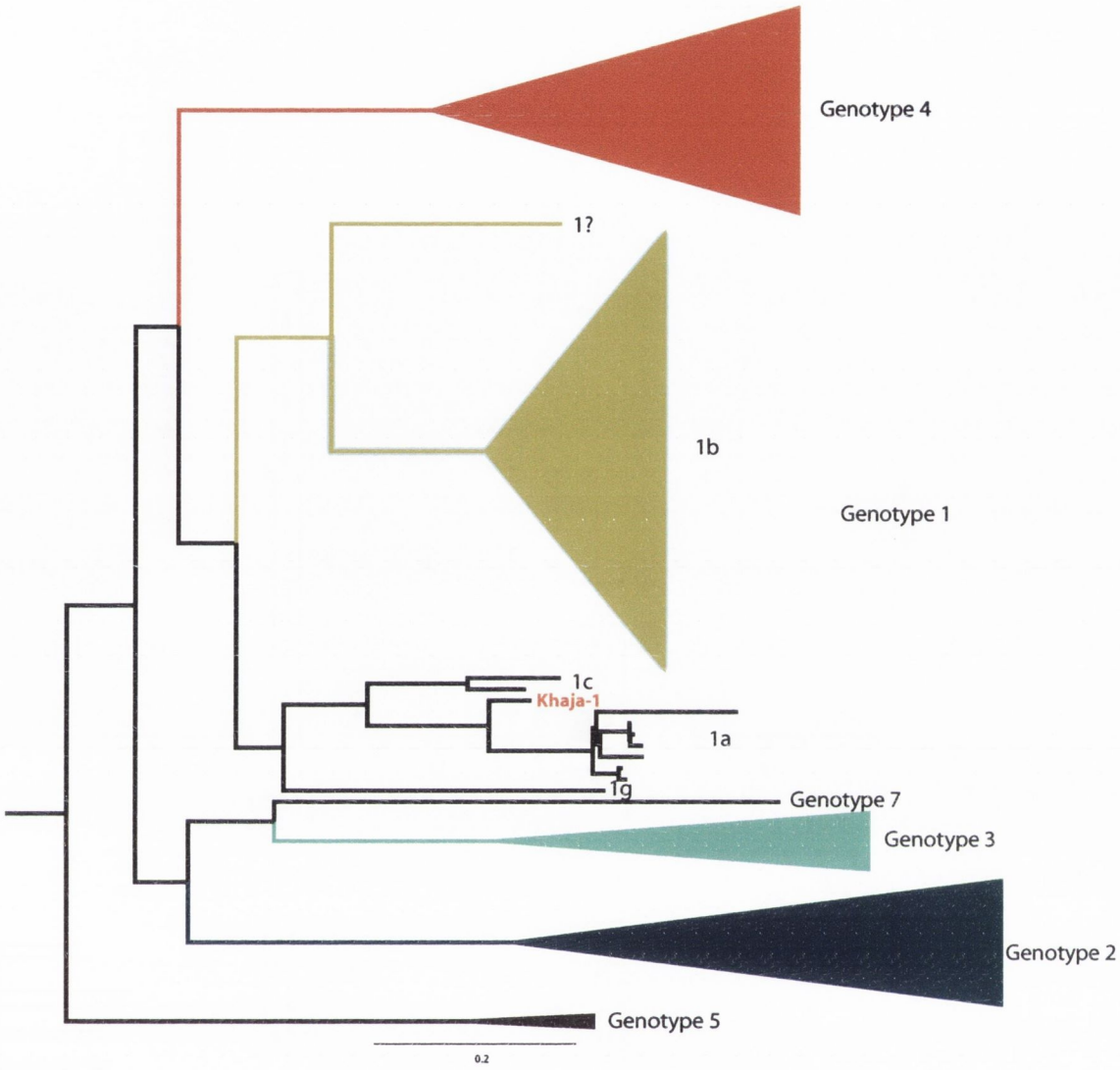


Figure 6.6: Maximum likelihood phylogenetic tree of the E2 gene depicting the topology of HCV isolates. For clarity purposes genotype 6 is not shown as the main focus is on genotype 1 with its putative recombinant highlighted in red. Each clade is labelled accordingly with respect to its genotype.

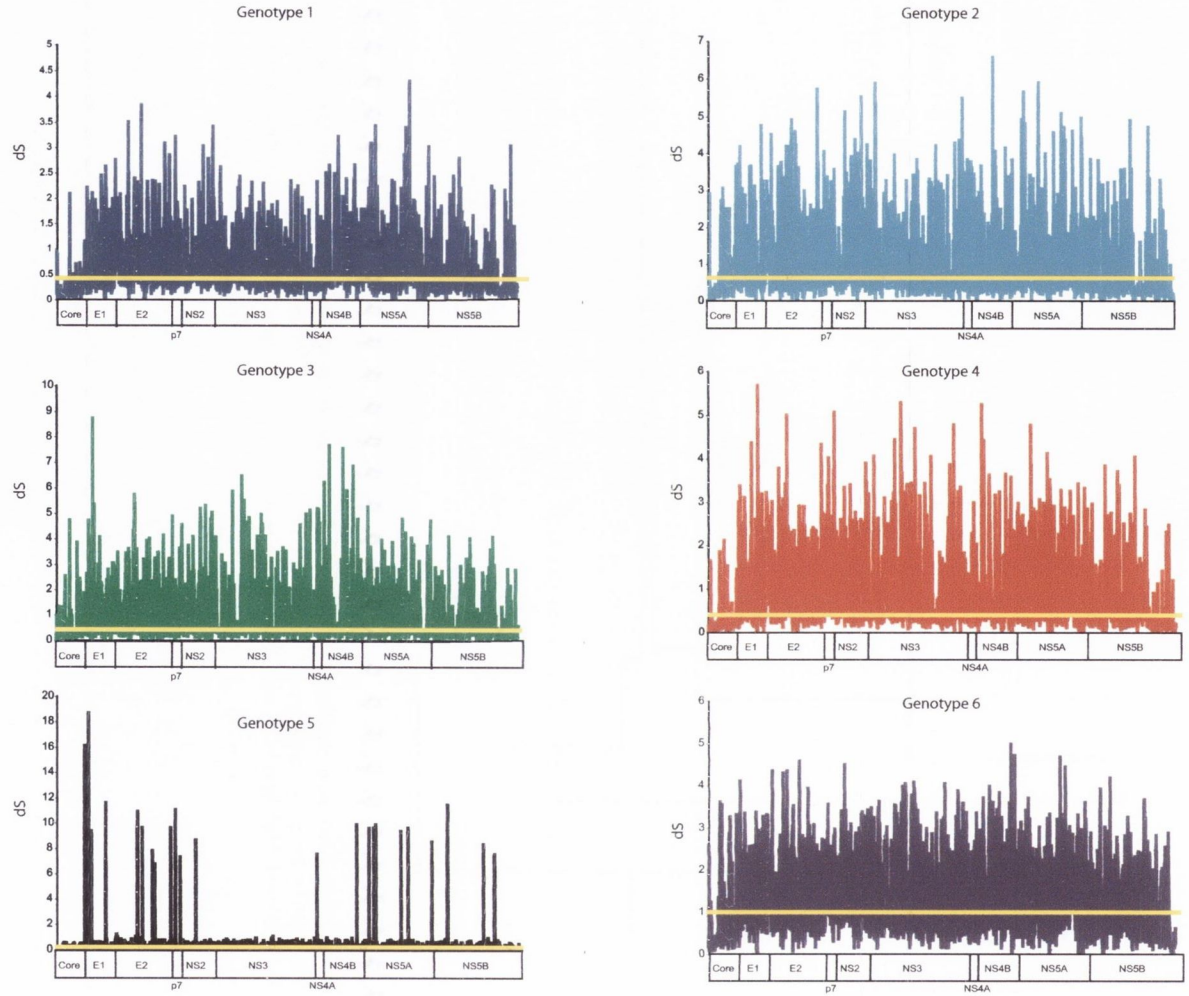


Figure 6.7: Distribution of synonymous nucleotide substitutions (d_S) in the polyprotein opening reading frame of HCV for all six genotypes. The medians for each genotype are show in yellow in all plots.

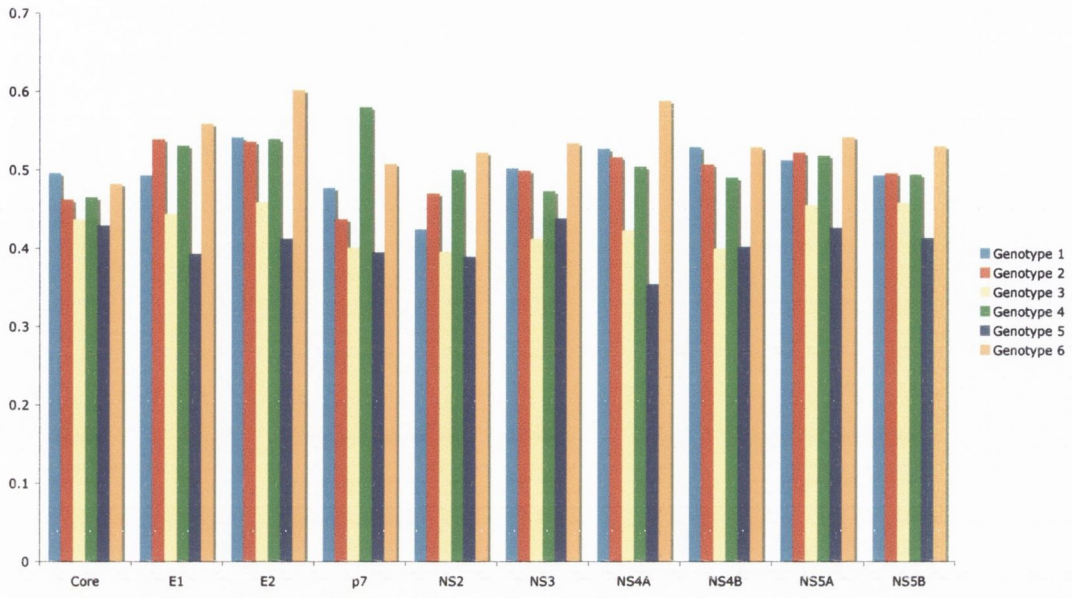


Figure 6.8: Codon adaptation index for each gene and its respective genotype

6.4.4 Codon adaptation index of genes and genotypes

Measuring the CAI has revealed that on average across all genes genotype 6 had higher CAI values than the other genotypes. Similarly, from a genomic point of view the E2 glycoprotein had on average higher CAI values followed by the phosphorylated NS5A protein and the E1 envelope protein (Figure 6.8).

6.4.5 Heterogeneity of SDMs

Using our method to search for slightly deleterious mutations we find a statistically significant abundance of SDMs in non structural proteins compared to structural proteins (t_2 = -3.596, P = 0.011) (Figure 6.9). In particular the RNA-dependent RNA polymerase NS5B contains the highest number for any genotype across the entire genome. In addition NS3 contains an abundance of SDMs in comparison to the remaining proteins encoded within the genome. In stark contrast the structural proteins which constitute the virion present a lower amount of SDMs with E2 containing the highest number. As well as genes presenting significant differences, genotypes also appear to undergo a heterogeneity of SDMs with genotype 6 containing the highest number of SDMs for nearly all genes.

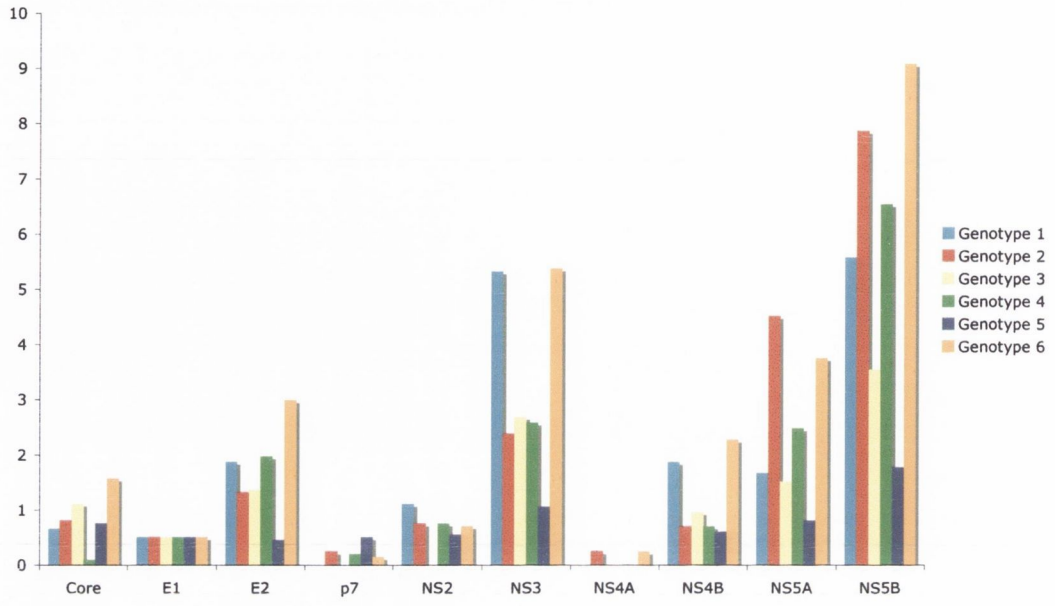


Figure 6.9: Comparison of the percentage of slightly deleterious mutations between HCV genes and genotypes. The percentages of SDMs are normalised along the genome and phylogeny and are therefore comparable between genotypes, between proteins or even between temporal sampling in the phylogeny (ancestral versus terminal distributions).

6.5 Discussion

Understanding the constraints that shape HCV evolution may reveal new avenues for antiviral therapeutics to curb the devastating infection caused by the virus. Previous studies focusing on the evolutionary dynamics of HCV have been limited for a number of reasons. Firstly, almost all analyses have been directed on genotype 1 and more specifically subtype 1b due to its worldwide dissemination. Secondly, typically a low sample size of isolates is used despite repeated evidence showing that methods for detecting selection are strongly influenced by sample size. Lastly, many of the above studies use simplistic measures of viral diversity and hence evaluation of what evolutionary processes are contributing on a genomic level remain unclear.

This study assembles a comprehensive set of full-length genomes and dissects the evolutionary constraints acting on each genotype for structural and non-structural proteins. We reveal for the first time the abundance of slightly deleterious mutations in the non-structural end of the ORF. From the CAI we predicted that on average the envelope glycoproteins E1 and E2 are higher expressed (Average 0.49 and 0.51 respectively) than any of the remaining proteins probably as a result of their presumed role in evading the host immune response. Therefore these genes are highly adaptable to their hosts where they are primarily responsible for the binding and entry of the virus to target cells. In addition we quantify the roles that recombination and adaptive evolution play in generating genetic diversity and how they can provide insights into

the epidemiological dynamics of HCV. The dichotomy between structural and non structural proteins may be as a result of natural selection forces although analysis of d_S indicates a highly homogeneous distribution across the genome. Taking all this information together it is clear that stochastic processes are important in shaping the genetic structure of HCV but we can not rule out immune-driven positive selection. These bouts of diversifying selection occur in the envelope glycoproteins and the nucleocapsid as witnessed from our results. Furthermore this study confirms that HCV rarely recombines.

6.5.1 Natural selection as the architect of HCV genotype evolution?

The elucidation of lineage-specific selective pressures in HCV serves as a useful tool to inform future vaccine design for genotypes. The role that natural selection plays in driving HCV diversity has never been fully subjected to rigorous examination for all genotypes. Therefore in an attempt to quantify the action of positive selection and negative selection on genotypes we utilised a maximum likelihood approach. For most genes one might expect positive selection to affect only a few amino acid residues along particular lineages. One attempt to detect signals of such local episodic natural selection is to use the branch site specific models (Zhang et al., 2005; Yang et al., 2005; Yang and Nielsen, 2002). Once an LRT suggests the presence of sites under positive selection, the empirical Bayes (EB) approach can be used to calculate the posterior probability. Sites with high posterior probabilities ($P > 95\%$) coming from the class with $\omega > 1$ are most likely under positive selection. With this in mind two approaches can be used to infer codons under positive selection in the foregrounds branches known as naive empirical Bayes (NEB) and the BEB. As previously mentioned in the results the BEB failed to identify a single site under positive selection while the NEB did infer sites. However, simulation studies suggest that NEB have high false-positive rates and was found to be unreliable in small data sets in site-based analysis as well. Hence, it is therefore considered unusable. This deficiency brought about the more reliable approach implemented by Yang et al. (2005), know as the BEB. In contrast, BEB appeared to be reliable, with the false-positive rate below 10% or 5% when sites were identified at the 90% or 95% cutoff (Yang et al., 2005). However, the power of the analysis can be very low. In this analysis the branch-site test of positive selection has provided significant support for the presence of positively selected sites on the foreground lineages, whereas no sites show a posterior probability higher than 95% according to BEB. Such a discrepancy is due to lack of power/information for any single site to reach high levels for the BEB probability to be strong enough. This may happen especially if positive selection has affected only a few lineages on the tree and is the case here. Reports of such observations are said to be common with simulations (Zhang et al., 2005; Yang, 2006). Nevertheless, our primary aim is to identify lineages not sites under positive selection as HCV proteins are poorly crystallised meaning that

sites cannot be reliably inferred to be structurally or functional important.

6.5.1.1 Previous Studies highlighting Adaptive Evolution in HCV

Previous studies have focused on exploring adaptive evolution within an individual HCV subtype most notably subtype 1b owing to its its clinical importance and large number of sequence data deposited in databases. Studies to date have been limited to the envelope glycoproteins E1 and E2 until recently when a detailed study entailing the selective pressures of the complete genome of HCV 1b isolates was performed (Campo et al., 2008). Brown et al. (2005) cloned HCV E1E2 full-length nucleotide sequences generated from serum samples of 4 chronically infected patients and identified 11 amino acids sites undergoing patient-specific adaptive evolution within the ectodomain of the E2 protein. Three of the four patients' selected sites mapped to a region known as the hypervariable region 1 (HVR1). This region is surface exposed (Yagnik et al., 2000) and has been proposed as a major target of the immune response possibly because its hypervariability is correlated with immune evasion (Weiner et al., 1992; Farci et al., 1996; Zibert et al., 1997). HVR1 is also known to contain potent, strain specific neutralizing antibody determinants (Farci et al., 1994, 1996; Shimizu et al., 1994). It is therefore well conceived that mutations occurring within the HVR1 leads to immune escape and subsequent development of chronic infection. The role that selection plays in the acute phase of HCV infection has also been investigated from the analyses of partial E1E2 sequences (Sheridan et al., 2004). This study highlighted a significant correlation between selective pressure and disease outcome with the fewest positively selected sites in fulminant HCV cases and the greatest number of selected sites in rapid progressors. This study also demonstrated that sites outside the HVR1 appear to play a major role in viral evolution and pathogenesis. More recently a full-length E1E2 clone panel of 45 gene sequences corresponding to genotypes that were previously under-represented was generated (Brown et al., 2007). This study confirmed that downstream of HVR1 a differential genotype-specific distribution of adaptive mutations was observed suggesting that subtly different selective pressures help to shape present-day genotype diversity. This sparks the idea that the heterogeneity of selective forces seen in genotypes promotes the current diversity and possible epidemiology seen in geographically distinct viral isolates. Although a powerful study at its time would clearly discount such a hypothesis and cite that no confounding factor appears to vary in a subtype-specific manner (Pybus et al., 2001), thus any differences in epidemic behaviour among HCV genotypes/subtypes are largely due to subtype-specific transmission patterns. An insightful study recently published (Campo et al., 2008) explored how coordinated variation among genomic sites under different forces of natural selection acts on full genome sequences belonging to the 1b subtype. Such coordination of substitutions in the polyprotein can be organised into a network that shares properties with biological, technological and social

complex networks. The detection of selection revealed that structural proteins have a high level of positive selection and are located towards the periphery of the network, while nonstructural proteins contain negatively selected sites of high centrality (Campo et al., 2008). Such evidence coincides with the findings here where positive selection is largely confined to structural proteins. Such confinement of adaptive substitutions to structural proteins may be responsible for fast adaptation to the host. On the other hand the severe restriction of non structural proteins by negative selection is probably due to functional or structural constraints where it is now emerging that *Flaviviridae* non-structural proteins are integral players in orchestrating virion morphogenesis (Murray et al., 2008). However, precise roles of HCV NS4B and NS5A proteins have as yet undefined role.

6.5.2 Recombination as a rare or absent process in HCV isolates

Recombination plays a significant role in the evolution of RNA viruses by creating genetic variation and its occurrence is widely documented in many families of RNA viruses (Aulicino et al., 2006; Boni et al., 2008; Chare et al., 2003; Chare and Holmes, 2006; Posada et al., 2002; Twiddy and Holmes, 2003; Worobey and Holmes, 1999). More important recombinants have been reported to occur now in a variety of members belonging to the family *Flaviviridae* such as Dengue virus (Aaskov et al., 2007; Holmes et al., 1999; Worobey et al., 1999; Vasilakis et al., 2007), St Louis encephalitis virus (Baillie et al., 2008), Japanese encephalitis virus (Twiddy and Holmes, 2003) and in pestiviruses, such as bovine viral diarrhoea virus (Jones and Weber, 2004). It has also been proposed that there is a difference in the detection of recombination with most of the mosquito-borne flaviviruses and none of the tick-borne flaviviruses undergoing recombination (Twiddy and Holmes, 2003). This difference may be accounted for by their differing modes of transmission reflected in biological and ecological factors of the particular host and vector species. In addition recombination breakpoints for nonsegemented positive-strand RNA viruses such as polioviruses and other picornaviruses are often located in regions of the genome encoding the first nonstructural protein but sometimes in structural proteins (mostly envelope proteins) (Kew et al., 2002; Guillot et al., 2000; Simmonds, 2006). Currently only a small number of naturally occurring recombinants have been identified for HCV despite frequent dual infection. Although natural inter-genotype recombinants (2k/1b) have been identified in St. Petersburg, Russia (Kalinina et al., 2002) and the discovery of recombinants appears to be speeding up with the existence of natural intragenotypic HCV recombinant strains (1a/1b) circulating in the Peruvian population (Colina et al., 2004), Vietnam (2i/6p) (Nopporpanth et al., 2006), the Philippines (2b/1b) (Kageyama et al., 2006) and more recently in France (2/5) (Legrand-Abravanel et al., 2007). The recombination sites of the previously described recombinant HCV strains are at the NS2/NS3 junction, the region of the genome that codes for first nonstructural proteins.

Therefore it is anticipated that the NS2/NS3 junction could be a favorable site for generating recombination events, especially with genotype 2. The same junction was also reported here to undergo recombination in 1a/1c isolates. Although no conflicting phylogenetic signal was seen within the non-structural genes such discrepancy arises probably due to the conservative nature of the test coupled with the diversity of the strains.

The search for recombination in this study revealed the existence of a 1a/1c intragenotypic recombinant with no recombination observed in any other genotypes. The finding of a 1a/1c intragenotypic recombinant strain was tentatively described before albeit with some doubt being placed on the precise location of the breakpoints and putative parental strains (Cristina and Colina, 2006). In their study a single method was used despite previous studies highlighting the inefficiency of using only a single method to detect recombination (Posada, 2002; Posada and Crandall, 2001). The method used was a sliding window method that slides a single window size along a sequence alignment and looks for differences in the support of phylogenetic tree within each window, a technique commonly employed in the study of viruses (Bull et al., 2007b; He et al., 2007; Salminen et al., 1995; Lole et al., 1999; Magiorkinis et al., 2005). Although the Simplot software has methodological limitations associated with it (Kurbanov et al., 2008; Simmonds and Midgley, 2005). Furthermore, Cristina and Colina (2006) used a large window size of 500 bp meaning that window sizes larger than the optimum could prevent the detection of small recombinant regions because the mean distance of that window will still give the correct pairwise sequence relationship while hiding the recombinant signal. In addition they failed to examine the effect of different sliding window sizes on producing discordant mosaic results (Zhang et al., 2008). Hence, this is why we choose a variety of different methods with varying sensitivities to examine discordant recombination patterns and breakpoint locations.

Interestingly from examination of the literature there is no apparent geographic proximity for the putative recombinant and parental strains. Thus, it is possible that the likely explanation of these apparent homologous recombination events is due to laboratory artifacts such as contamination or other experimental errors like template switching during amplification in a mixed sample and therefore this strain does not represent a natural recombinant.

It would appear that the true frequency of recombination in HCV is quite low and this is somewhat surprising given the genotype epidemiology with a wide range of genotypes co-circulating. In addition infection in these areas is often characterised by frequent repeated needle sharing with several infected individuals, a prequisite condition for recombination that principally, either vector or host must be co-infected with greater than one strain of the virus in question. Therefore HCV fulfills both epidemiological opportunity and biological compatibility. However, in most cases in positive sense RNA viruses such as this one, recombination appears to be a sporadic event that does not occur at high enough frequency to make it a key evolutionary strategy (Holmes and Drummond, 2007). If proved to occur it would facilitate two evolutionary

processes: the purging of deleterious mutations and the rapid generation of novel variants. The lack of evidence of recombination for HCV suggests that such events rare in vivo resulting in nonviable recombinants or that the true frequency of recombination of HCV is underestimated. For example a process where very short sections of RNA are transferred between viruses would be undetectable by the majority of other methods devised to detect recombination. Similarly, recombination would not be detected easily between variants of the same genotype and in highly diverse geographic areas inter-genotype recombinants would remain difficult to document due to the lack of a complete catalogue of variants. According to recent literature recombination should be considered as a potentially relevant mechanism generating genetic variation in HCV (Sentandreu et al., 2008). This detailed study focused for the first time at searching for recombination at the intrapatient intragenic level which is the lowest possible level of diversity. By analysing close to 18000 sequences isolated from 111 patients assigned to different clinical groups the authors assert the opinion that the frequency of recombination in HCV can be a much more common phenomenon than previously recognised.

6.5.3 Genotype diversity

We propose that much of the divergence among genotypes of HCV may be explained by drift in which genetic isolation facilitates accumulation and fixation of slightly deleterious (neutral) mutations. Time of isolation then accounts for the amount of divergence between strains. This is supported from the analysis of synonymous sites and SDMs where mutations that are slightly deleterious are fixed in the population by genetic drift when the effective population size is small. Viral populations generally present significant sizes within their hosts but they undergo strong bottlenecks during the infection of other host individuals or in the switch to different stages during the viral life cycle. This dynamic favours the fixation of slightly deleterious mutations. The higher rate of fixation of SDMs in non-structural compared to structural proteins as tested in our study lends support to the greater permissibility of non-structural proteins to accumulate mutations by genetic drift. SDMs may have accumulated as a result of either a founder effect, as a result of a selective process or due to a change in the selection-drift balance in each genotype. Because a greater effect of genetic drift would relax constraints in synonymous and non-synonymous sites, we should observe higher dS values at synonymous sites when genetic drift is more important. Our analyses of synonymous nucleotide substitutions clearly demonstrate that synonymous sites are much more relaxed in type 6 isolates than in other genotypes although no significant differences were observed when structural proteins were compared with non-structural proteins along the genome. The reason therefore for the increasing number of SDMs in type 6 may have been very likely the result of relaxed constraints in genotype 6 isolates.

Isolates of HCV genotype 6 have three unique features. First, they are endemic in Southeast

Asia or its surroundings. Second, they represent the most genetically diverse and complex HCV lineage. Third, an older origin of diversification than the other genotypes has been suggested (Salemi and Vandamme, 2002). The number of subtypes belonging to genotype 6 is now 21, the genetic diversity of which exceeds any of the other five HCV genotypes. Hence, this genotype appears to harbour the most evolutionary and epidemic potential in comparison to the rest where it is becoming increasingly detected among Asian immigrants in North America. Caution must be emphasised if these variants enter effective transmission networks such as IDUs where significant changes in prevalence and associated disease may be induced.

The genetic diversity that constitutes HCV is represented by seven genotypes, over 80 assigned subtypes and many additional unassigned variants and, very likely additional missing genotypes and subtypes. Therefore determining the dominating factor driving the global epidemiology of HCV is of enormous clinical importance with repeated studies showing that the response to antiviral therapy is lower in patients with subtypes 1a, 1b and 4a than in those infected with type 2 and type 3 strains. While it is clear that genetic drift is responsible for much of the variability observed between geographically isolated genotypes of HCV, the continuing persistence and transmission of HCV into human populations are in part facilitated by molecular adaptation from contributing evolutionary factors. Drift and selection are quantitatively not mutually exclusive forces and may occur concurrently. Thus, it is the combined effects of both forces acting simultaneously on different regions of the genome that ultimately determine the extent of divergence within viral species. Also the finding of a number of highly conserved genes under strong purifying selection is important for antibody based vaccine design. The likelihood of escape from these antibodies is low and therefore such antibodies targeted at regions of low amino acid variability could have great therapeutic potential. Furthermore it is vital to understand the degree of polymorphism and its response to antiviral treatment which until recently has not been extensively evaluated (Cuevas et al., 2008; Lopez-Labrador et al., 2008; Torres-Puente et al., 2008b,a). Altogether, these results suggest that viral strategies to evade selection pressure from the immune system and antiviral therapies do not result from a single mechanism but are based on a combination of changes along the genome.

6.6 Acknowledgements

Our thanks go to Alma Bracho for providing the genome of 1g and to Thomas Kuntzen for the early release of genotype 4 sequences.

Chapter 7

General discussion

Taken together the five studies above present evidence of an intimate relationship between changes in the evolutionary rates of encoded proteins and viral epidemiology. The single-gene study in Chapters 2 and 3 are primarily a demonstration of the interchangeability of serotype dynamics in the evolutionary history of a capsid gene (VP1) in a major animal virus (FMDV). This illustrates how viruses belonging to different parts of the world and even different hosts evolve in a serotype-specific manner. Five of the seven serotypes were detected to be under positive selection and comparison of the regions highlighted diverse patterns of selective constraints operating at the molecular and phylogenetic level. Indeed all those serotypes under positive selection comprise a similar region around N131- R149 that is located in the immunodominant site A within the G-H loop. Remarkably only serotypes A and O, the wider geographically spread FMDV serotypes, have the full RGD motif under positive selection. This motif is conserved across all serotypes and plays a direct role in the interaction with host cell receptors. The African serotypes showed considerable evidence of positive selection at overlapping VP1 protein regions. In fact, all of the regions detected in SAT 1 to be under adaptive evolution were also found in SAT 2 to be positively selected (Table 2.1). Most of the regions under positive selection in the SAT serotypes are not found in non-SAT types sparking speculation that there is a difference in the selective pressures. Such evolutionary differences are further enhanced from protein structure analysis of the VP1 gene where certain serotypes have accumulated compensatory mutations. This probably arose as a strategy to ameliorate the effect of slightly deleterious mutations fixed by genetic drift. A more recent study validated the vast majority of residues detected to be under selection consistent with their location in loop regions (Lewis-Rogers et al., 2008).

With such a large degree of sequence diversity observed within FMDV, it became evident to us that estimating the date of origin of the FMDV MRCA was vital. Knowing this date would provide a timescale in which the rapid rate of evolution within subtypes has occurred and

therefore, would provide a standard to which future rates of evolution could be predicated. No previous analysis has attempted to estimate the date of origin of FMDV and its serotypes due to the lack of sequences covering a broad range range of sampling dates. However our globally constructed dataset containing sequences of all serotypes spatially and temporally spread has counteracted this problem and we have successfully estimated the date of origin of the MRCA of FMDV. This analysis suggested a date in the middle of the 16th century for the origin of an ancient precursor virus that subsequently underwent two separate diversification events in a relatively short time frame to give rise to the the current day SAT type viruses and Euro-Asiatic type viruses. Also this timing coincides with the first historical sighting of the disease in Northern Italy and is further validated by the time-line of European exploration. The timing of mass European exploration to foreign soils has undoubtedly played a role in the spread of the disease and in fact we speculate that such events may have caused the transmission of the disease from Europe into Africa. Another reason why this type of analysis has never been carried out before is largely because until recently accounting for rate variation among lineages was not possible in a phylogenetic framework (Drummond et al., 2006). Thus, the development of relaxed clock models means that our dating times are much more realistic as several studies have indicated that RNA viruses violate the traditional molecular clock hypothesis (Pybus et al., 2001; Salemi and Vandamme, 2002; Pybus et al., 2005; Aulicino et al., 2006; Bryant et al., 2007; Davis et al., 2007; Zhou and Holmes, 2007; Duffy et al., 2008; Hughes, 2008).

The genomic complexity of FMDV was unveiled in Chapter 4 where complete genomes for all serotypes were examined. Our results indicate that diversifying selection has driven the evolution of FMDV genomes. A positive correlation between the absolute proportion of codons under adaptive evolution and the amount of exposed surface is observed that reflects the evolutionary dynamic nature of capsid proteins in relation to its ability to interact with host cells. From the analysis of synonymous substitutions it is clear that they are heterogeneously enhanced in SAT serotypes compared to Euro-Asiatic serotypes in particular in structural proteins. Indeed we conclude that relaxed constraints in structurally encoded exposed proteins may facilitate the emergence of new polymorphic immune-escaping mutants in the population. These relaxed constraints may have permitted the accumulation of slightly deleterious mutations that once compensated for by the subsequent fixation of conditional-advantageous mutations in nearby structural regions, became advantageous to generate immune-escape mutants. In fact this is the case with SAT serotypes that contain a high proportion of compensatory mutations in structural proteins in ancestral lineages.

The single gene study and genome wide study of FMDV provides valuable information for any future studies into the evolution and development of FMDV. The complex evolutionary dynamic between serotypes is displayed which exacerbates the matter of vaccination and purges the notion of a universal antiviral strategy.

The second part of the thesis is confined to the study of two major pathogenic human viruses namely HIV-1 and HCV. Chapter 5 deals with dissecting the nature of coevolution on two functionally important genes. Firstly, the ability of HIV-1 to adapt to humans is thought to have required a single amino acid change (Wain et al., 2007) but yet amino acids rarely function completely independent of each other. The nature of intra-molecular dependency between amino acids can now be investigated using robust statistical tools. Therefore this served as a useful example to demonstrate if indeed a single amino acid is responsible for the adaptation of HIV-1 or does a network of residues acting in concert to perform a specific function better explain the event. Our results indicate that this single amino acid is highly reliant on its interaction with other amino acids in functional domains probably to carry out a specific function. The function in question in the case is the adaptation of HIV-1 from simians to human hosts. Although experimental evidence implicated a single amino acid responsible we firmly believe that from an evolutionary perspective this site did not act alone and was aided by surrounding amino acids. The next part of this chapter dealt with identifying those functional or structurally important residues whose evolution is constrained over the entire group M pandemic in the functionally important envelope gene. Using the same approach as before we predicted a number of pairs of coevolving residues whose dependency is mediated by specific function such as CD4 binding, coreceptor binding and glycosylation. Those pairs not assigned biological significance may represent novel targets for site directed mutagenesis to aid in the identification of functional domains or even serve as a strategy for the design of broadly neutralising vaccines. Future work could address subtype-specific coevolving mechanisms as these may provide additional clues into subtype-specific mechanisms of infectivity.

The main goal of the HCV study was to address some of the questions concerning the evolutionary patterns of different genotypes. For this reason, we analysed 133 full-genome sequences representative of the six major phylogenetic clades of HCV. We addressed the significance that recombination and adaptive evolution play in driving the epidemic and sparking new emerging subtypes. Our results lead us to the conclusion that recombination as previously conceived is a rare and absent process in viral isolates. Similarly adaptive evolution appears in short bursts at the molecular level and doesn't seem to account for observed sequence diversity at the population level. Then the predominant mode of evolution acting on HCV is genetic drift with which analysis of synonymous sites and slightly deleterious mutations both concur. This, coupled with varying transmission routes, probably serves as an efficient model to spread. Further studies, and a detailed knowledge of the evolutionary rates in the different lineages may serve to be instrumental in addressing the origin and evolution of the virus. Thus, it appears that stochastic processes of viral migration and adaptive evolution play a vital role in shaping short-term evolutionary dynamics with future patterns of outbreaks inherently complex and requiring intensive surveillance.

Irrespective of the RNA virus under study and its ability to infect animals, humans or plants there is no strict or inherent prerequisite pattern of evolution. In this work it is apparent that there are a multitude of factors that need to be accounted for before a coherent explanation of viral behaviour is made. However, what is clear is that that RNA viruses contain a significant amount of plasticity from which they rapidly mutate and spread into new susceptible hosts subject to successful adaptation. Therefore it would appear all RNA viruses harbour a similar common mechanisms for evolutionary change but yet few exploit this to an unprecedented level. In particular a variety of different phylogeographic patterns are observed in human viruses reflecting their rates of viral gene flow, mode of transmission and time of association with humans populations (Holmes, 2004, 2008). For example, HCV displays relatively strong spatial subdivision where genetic diversity is partitioned into a set of geographical distinct distributed genotypes and subtypes and often related with particular risk groups (Simmonds, 2004; Pybus et al., 2005). Conversely, wavelike transmission has been proposed for the transmission of Ebola virus from Zaire since 1976 (Walsh et al., 2005). While influenza virus represents a complex level of diversity and follows a sink-source model in which new lineages are seeded from a persistent influenza reservoir located in the tropical regions, to sink populations in temperate regions (Rambaut et al., 2008). The extent and nature of ecological interactions among pathogens still remain to be fully understood. Yet some studies have implicated ecological rather than genetic factors as the main determinant of successful emergence as with Venezuelan equine encephalitis virus (VEEV) (Anishchenko et al., 2006) and the pandemic variant of HIV/AIDS (Gilbert et al., 2007). Nevertheless the complex interplay between ecology and genetics is likely to play a key role in the emergence of viruses.

Paradoxically, the beauty behind RNA viruses is it allows us to perform a simple link between experimental and natural systems (Elena et al., 2008). This trait allows us to determine if viral evolution in nature is similar to that established *in vitro* - a feature inherently present in few organisms. Inessence viruses represent a delicate group of pathogens modulated by host immunity and surrounded by evolutionary complexity. Their stealth is derived from their small compact genomes as evolving larger complex genomes might be difficult for RNA viruses because they have to evolve mechanisms to allow additional genes to be translated. Future studies directed on viral evolution need careful attention as we are on the cusp on a genomic sequencing renaissance where cheap mass scale sequencing known as pyrosequencing technology offers the imminent release of 100,000s of sequences (Margulies et al., 2005; Eriksson et al., 2008). This advance in genomic sequencing creates a major intellectual challenge with insufficient computational tools to analyse such enormous data. In-depth analysis of many thousands of viral genomes is now on the horizon in addressing key questions of biological importance but yet the power of genomic data is rarely fully appreciated. As such it is imperative that the standard analysis of viral genome data be coupled with detailed functional, experimental and epidemiological information

to give a true value for the power of genomic data. This has promoted some researchers to move towards an automating approach from recent developments in eScience specifically a "workflow" approach. Workflows are a formalisation of traditional bioinformatics analysis, where the process of data download, processing and analysis is described and carried out within an appropriate computing environment. One such system that shows future promise is the myGrid project although its application to viral data remains unperformed (Stevens et al., 2003, 2004). In this case workflows are designed and run using tools implemented in the Taverna workflow system which has provided software to the bioinformatics community allowing scientists to access many functions traditionally associated with grid computing. These functions include job execution, coordination in the form of dataflow workflows and management of results, virtual experiments. One such advantage of shifting beyond the use of individual software programs to state of the art workflows is that all aspects of a particular analysis is recorded. This explicit recording of the analytical process in addition to the results mean that workflows can be published and reused, allowing the replication of our analyses both as the project progresses and the databases grow (Stevens et al., 2004). Thus, such approaches will undoubtedly pioneer the way in which the wealth of evolutionary and epidemiological data contained within viral genomes is analysed. Yet researchers must strive to account for the structural constraints associated with RNA viruses and attempt to account or model for recombination when designing new statistical tools to test for example the process of adaptive evolution.

Other biological problems associated with viruses constrain our understanding of them. For instance with HCV the lack of a cheap efficient animal model has prevented a considerable amount of progress from been made in eradicating the disease. In addition the experimental evaluation of subtype and genotype fitness needs to performed. This is critical to our development of subtype-specific based vaccines. Other challenges remaining ahead include the nature of epistatic interactions among genes and their role in the development of drug resistance and immune escape.

Bibliography

Aaskov, J., Buzacott, K., Field, E., Lowry, K., Berlioz-Arthaud, A., and Holmes, E. C., 2007. Multiple recombinant dengue type 1 viruses in an isolate from a dengue patient. *J Gen Virol*, **88**(Pt 12):3334–3340.

Abebe, A., Demissie, D., Goudsmit, J., Brouwer, M., Kuiken, C. L., Pollakis, G., Schuitemaker, H., Fontanet, A. L., and Rinke de Wit, T. F., 1999. Hiv-1 subtype c syncytium- and non-syncytium-inducing phenotypes and coreceptor usage among ethiopian patients with aids. *AIDS*, **13**(11):1305–1311.

Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D., and Brown, F., 1989. The three-dimensional structure of foot-and-mouth disease virus at 2.9 a resolution. *Nature*, **337**(6209):709–16–.

Aggarwal, N. and Barnett, P., 2002. Antigenic sites of foot-and-mouth disease virus (fmdv): an analysis of the specificities of anti-fmdv antibodies after vaccination of naturally susceptible host species. *J Gen Virol*, **83**(Pt 4):775–82–.

Akashi, H., 1999. Within- and between-species dna sequence variation and the 'footprint' of natural selection. *Gene*, **238**(1):39–51.

Alberti, A., Chemello, L., and Benvegnu, L., 1999. Natural history of hepatitis c. *J Hepatol*, **31 Suppl 1**:17–24.

Ane, C., Burleigh, J. G., McMahon, M. M., and Sanderson, M. J., 2005. Covarion structure in plastid genome evolution: a new statistical test. *Mol Biol Evol*, **22**(4):914–924.

Anishchenko, M., Bowen, R. A., Paessler, S., Austgen, L., Greene, I. P., and Weaver, S. C., 2006. Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. *Proc Natl Acad Sci U S A*, **103**(13):4994–4999.

Anisimova, M. and Liberles, D. A., 2007. The quest for natural selection in the age of comparative genomics. *Heredity*, **99**(6):567–579.

Anisimova, M., Nielsen, R., and Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**(3):1229–1236.

Ansell, D. M., Samuel, A. R., Carpenter, W. C., and Knowles, N. J., 1994. Genetic relationships between foot-and-mouth disease type asia 1 viruses. *Epidemiol Infect*, **112**(1):213–224.

Araujo, J. P. J., Montassier, H. J., and Pinto, A. A., 2002. Extensive antigenic and genetic variation among foot-and-mouth disease type a viruses isolated from the 1994 and 1995 foci in sao paulo, brazil. *Vet Microbiol*, **84**(1-2):15–27.

Arbiza, L., Dopazo, J., and Dopazo, H., 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol*, **2**(4):e38.

Arien, K., Vanham, G., and Arts, E., 2007. Is hiv-1 evolving to a less virulent form in humans? *Nat Rev Microbiol*, **5**(2):141–51–.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W., 2000. Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Mol Biol Evol*, **17**(1):164–178.

Aulicino, P., Holmes, E., Rocco, C., Mangano, A., and Sen, L., 2006. Extremely rapid spread of hiv-1 bf recombinants in argentina. *J Virol*, :–.

Awadalla, P., 2003. The evolutionary genomics of pathogen recombination. *Nat Rev Genet*, **4**(1):50–60.

Baillie, G. J., Kolokotronis, S.-O., Waltari, E., Maffei, J. G., Kramer, L. D., and Perkins, S. L., 2008. Phylogenetic and evolutionary analyses of st. louis encephalitis virus genomes. *Mol Phylogenet Evol*, **47**(2):717–728.

Bakewell, M. A., Shi, P., and Zhang, J., 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A*, **104**(18):7489–7494.

Baldwin, C. E. and Berkhout, B., 2006. Second site escape of a t20-dependent hiv-1 variant by a single amino acid change in the cd4 binding region of the envelope glycoprotein. *Retrovirology*, **3**:84.

Ball, S. C., Abraha, A., Collins, K. R., Marozsan, A. J., Baird, H., Quinones-Mateu, M. E., Penn-Nicholson, A., Murray, M., Richard, N., Lobritz, M., *et al.*, 2003. Comparing the ex vivo fitness of ccr5-tropic human immunodeficiency virus type 1 isolates of subtypes b and c. *J Virol*, **77**(2):1021–1038.

Baltimore, D., 1971. Expression of animal virus genomes. *Bacteriol Rev*, **35**(3):235–241.

Baranowski, E., Ruiz-Jarabo, C., Lim, F., and Domingo, E., 2001. Foot-and-mouth disease virus lacking the vp1 g-h loop: the mutant spectrum uncovers interactions among antigenic sites for fitness gain. *Virology*, **288**(2):192–202–.

Barton, N. H. and Charlesworth, B., 1998. Why sex and recombination? *Science*, **281**(5385):1986–1990.

Bastos, A., Haydon, D., Forsberg, R., Knowles, N., Anderson, E., Bengis, R., Nel, L., and Thomson, G., 2001. Genetic heterogeneity of sat-1 type foot-and-mouth disease viruses in southern africa. *Arch Virol*, **146**(8):1537–51–.

Bastos, A., Haydon, D., Sangare, O., Boshoff, C., Edrich, J., and Thomson, G., 2003. The implications of virus diversity within the sat 2 serotype for control of foot-and-mouth disease in sub-saharan africa. *J Gen Virol*, **84**(Pt 6):1595–606–.

Bastos, A. D., Boshoff, C. I., Keet, D. F., Bengis, R. G., and Thomson, G. R., 2000. Natural transmission of foot-and-mouth disease virus between african buffalo (syncerus caffer) and impala (aepyceros melampus) in the kruger national park, south africa. *Epidemiol Infect*, **124**(3):591–598.

Batschelet, E., Domingo, E., and Weissmann, C., 1976. The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate. *Gene*, **1**(1):27–32.

Beerli, P. and Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A*, **98**(8):4563–4568.

Belsham, G. J., 1993. Distinctive features of foot-and-mouth disease virus, a member of the picornavirus family; aspects of virus protein synthesis, protein processing and structure. *Prog Biophys Mol Biol*, **60**(3):241–260.

Belshaw, R., Gardner, A., Rambaut, A., and Pybus, O. G., 2008. Pacing a small cage: mutation and rna viruses. *Trends Ecol Evol*, **23**(4):188–193.

Belshaw, R., Pybus, O. G., and Rambaut, A., 2007. The evolution of genome compression and genomic novelty in rna viruses. *Genome Res*, **17**(10):1496–1504.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R., and Ben-Tal, N., 2004. Conseq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**(8):1322–1324.

Berglund, A.-C., Wallner, B., Elofsson, A., and Liberles, D. A., 2005. Tertiary windowing to detect positive diversifying selection. *J Mol Evol*, **60**(4):499–504.

Bernstein, H. B., Tucker, S. P., Kar, S. R., McPherson, S. A., McPherson, D. T., Dubay, J. W., Lebowitz, J., Compans, R. W., and Hunter, E., 1995. Oligomerization of the hydrophobic heptad repeat of gp41. *J Virol*, **69**(5):2745–2750.

Bickel, P. J., Cosman, P. C., Olshen, R. A., Spector, P. C., Rodrigo, A. G., and Mullins, J. I., 1996. Covariability of v3 loop amino acids. *AIDS Res Hum Retroviruses*, **12**(15):1401–1411.

Blaser, M. J., 1997. Heterogeneity of helicobacter pylori. *Eur J Gastroenterol Hepatol*, **9 Suppl 1**:S3–6; discussion S6–7.

Blok, J., Mackenzie, A., Guy, P., and Gibbs, A., 1987. Nucleotide sequence comparisons of turnip yellow mosaic virus isolates from australia and europe. *Arch Virol*, **97**(3-4):283–295.

Boge, M., Wyss, S., Bonifacino, J. S., and Thali, M., 1998. A membrane-proximal tyrosine-based signal mediates internalization of the hiv-1 envelope glycoprotein via interaction with the ap-2 clathrin adaptor. *J Biol Chem*, **273**(25):15773–15778.

Bolser, D., Dafas, P., Harrington, R., Park, J., and Schroeder, M., 2003. Visualisation and graph-theoretic analysis of a large-scale protein structural interactome. *BMC Bioinformatics*, **4**:45.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., and Petropoulos, C. J., 2004. Evidence for positive epistasis in hiv-1. *Science*, **306**(5701):1547–1550.

Boni, M. F., Posada, D., and Feldman, M. W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, **176**(2):1035–1047.

Boni, M. F., Zhou, Y., Taubenberger, J. K., and Holmes, E. C., 2008. Homologous recombination is very rare or absent in human influenza a virus. *J Virol*, **82**(10):4807–4811.

Bracho, M. A., Carrillo-Cruz, F. Y., Ortega, E., Moya, A., and Gonzalez-Candelas, F., 2006a. A new subtype of hepatitis c virus genotype 1: complete genome and phylogenetic relationships of an equatorial guinea isolate. *J Gen Virol*, **87**(Pt 6):1697–1702.

Bracho, M. A., Gosalbes, M. J., Gonzalez, F., Moya, A., and Gonzalez-Candelas, F., 2006b. Molecular epidemiology and evolution in an outbreak of fulminant hepatitis b virus. *J Clin Microbiol*, **44**(4):1288–1294.

Bracho, M. A., Saludes, V., Martro, E., Bargallo, A., Gonzalez-Candelas, F., and Ausina, V., 2008. Complete genome of a european hepatitis c virus subtype 1g isolate: phylogenetic and genetic analyses. *Virol J*, **5**:72.

Brooksby, J. B., 1958. The virus of foot-and-mouth disease. *Adv Virus Res*, **5**:1–37.

Brooksby, J. B. and Rogers, J., 1957. Methods used in typing the virus of foot-and-mouth disease at pirbright, 1950-55. Technical Report 31-34, European Productivity Agency of the Organisation for European Economic Co-operation, Project No208, Paris.

Brown, C. J., Garner, E. C., Dunker, A. K., and Joyce, P., 2001. The power to detect recombination using the coalescent. *Mol Biol Evol*, **18**(7):1421–1424.

Brown, R. J. P., Juttla, V. S., Tarr, A. W., Finnis, R., Irving, W. L., Hemsley, S., Flower, D. R., Borrow, P., and Ball, J. K., 2005. Evolutionary dynamics of hepatitis c virus envelope genes during chronic infection. *J Gen Virol*, **86**(Pt 7):1931–1942.

Brown, R. J. P., Tarr, A. W., McClure, C. P., Juttla, V. S., Tagiuri, N., Irving, W. L., and Ball, J. K., 2007. Cross-genotype characterization of genetic diversity and molecular adaptation in hepatitis c virus envelope glycoprotein genes. *J Gen Virol*, **88**(Pt 2):458–469.

Bryant, J., Holmes, E., and Barrett, A., 2007. Out of africa: a molecular perspective on the introduction of yellow fever virus into the americas. *PLoS Pathog*, **3**(5):e75–.

Bull, J. J., Sanjuan, R., and Wilke, C. O., 2007a. Theory of lethal mutagenesis for viruses. *J Virol*, **81**(6):2930–2939.

Bull, R. A., Tanaka, M. M., and White, P. A., 2007b. Norovirus recombination. *J Gen Virol*, **88**(Pt 12):3347–3359.

Burch, C. L. and Chao, L., 2004. Epistasis and its relationship to canalization in the rna virus phi 6. *Genetics*, **167**(2):559–567.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., *et al.*, 2005. Natural selection on protein-coding genes in the human genome. *Nature*, **437**(7062):1153–1157.

Caffrey, M., Cai, M., Kaufman, J., Stahl, S. J., Wingfield, P. T., Covell, D. G., Gronenborn, A. M., and Clore, G. M., 1998. Three-dimensional solution structure of the 44 kda ectodomain of siv gp41. *EMBO J*, **17**(16):4572–4584.

Campo, D. S., Dimitrova, Z., Mitchell, R. J., Lara, J., and Khudyakov, Y., 2008. Coordinated evolution of the hepatitis c virus. *Proc Natl Acad Sci U S A*, **105**(28):9685–9690.

Carrillo, C., Lu, Z., Borca, M. V., Vagnozzi, A., Kutish, G. F., and Rock, D. L., 2007. Genetic and phenotypic variation of foot-and-mouth disease virus during serial passages in a natural host. *J Virol*, **81**(20):11341–11351.

Carrillo, C., Tulman, E., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., Kutish, G., and Rock, D., 2005. Comparative genomics of foot-and-mouth disease virus. *J Virol*, **79**(10):6487–504–.

Carrington, C., Foster, J., Pybus, O., Bennett, S., and Holmes, E., 2005. Invasion and maintenance of dengue virus type 2 and type 4 in the americas. *J Virol*, **79**(23):14680–7–.

Chamary, J. V., Parmley, J. L., and Hurst, L. D., 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, **7**(2):98–108.

Chao, L., 1990. Fitness of rna virus decreased by muller's ratchet. *Nature*, **348**(6300):454–455.

Chare, E. R., Gould, E. A., and Holmes, E. C., 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense rna viruses. *J Gen Virol*, **84**(Pt 10):2691–2703.

Chare, E. R. and Holmes, E. C., 2006. A phylogenetic survey of recombination frequency in plant rna viruses. *Arch Virol*, **151**(5):933–946.

Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H., and Benner, S. A., 1997. An analysis of simultaneous variation in protein structures. *Protein Eng*, **10**(4):307–316.

Chen, B., Vogan, E. M., Gong, H., Skehel, J. J., Wiley, D. C., and Harrison, S. C., 2005. Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature*, **433**(7028):834–841.

Choisy, M., Woelk, C., Guegan, J., and Robertson, D., 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol*, **78**(4):1962–70–.

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W., and Houghton, M., 1989. Isolation of a cdna clone derived from a blood-borne non-a, non-b viral hepatitis genome. *Science*, **244**(4902):359–362.

Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W. C., Moore, J. P., Trkola, A., and Morris, L., 2003. The ccr5 and cxcr4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype c. *J Virol*, **77**(7):4449–4456.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., *et al.*, 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**(5652):1960–1963.

Cleaveland, S., Haydon, D. T., and Taylor, L., 2007. Overviews of pathogen emergence: which pathogens emerge, when and why? *Curr Top Microbiol Immunol*, **315**:85–111.

Cleaveland, S., Laurenson, M. K., and Taylor, L. H., 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond B Biol Sci*, **356**(1411):991–999.

Clements, M. L., Subbarao, E. K., Fries, L. F., Karron, R. A., London, W. T., and Murphy, B. R., 1992. Use of single-gene reassortant viruses to study the role of avian influenza a virus genes in attenuation of wild-type human influenza a virus for squirrel monkeys and adult human volunteers. *J Clin Microbiol*, **30**(3):655–662.

Codoner, F. M., Daros, J.-A., Sole, R. V., and Elena, S. F., 2006a. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathog*, **2**(12):e136.

Codoner, F. M. and Elena, S. F., 2008. The promiscuous evolutionary history of the family bromoviridae. *J Gen Virol*, **89**(Pt 7):1739–1747.

Codoner, F. M. and Fares, M., 2008. Why should we care about molecular coevolution. *Evolutionary Bioinformatics*, **4**(237-246).

Codoner, F. M., Fares, M. A., and Elena, S. F., 2006b. Adaptive covariation between the coat and movement proteins of prunus necrotic ringspot virus. *J Virol*, **80**(12):5833–5840.

Codoner, F. M., O'Dea, S., and Fares, M. A., 2008. Reducing the false positive rate in the non-parametric analysis of molecular coevolution. *BMC Evol Biol*, **8**:106.

Colina, R., Casane, D., Vasquez, S., Garcia-Aguirre, L., Chunga, A., Romero, H., Khan, B., and Cristina, J., 2004. Evidence of intratypic recombination in natural populations of hepatitis c virus. *J Gen Virol*, **85**(Pt 1):31–37.

Comeron, J. M., 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol*, **41**(6):1152–1159.

Condy, J. B., Hedger, R. S., Hamblin, C., and Barnett, I. T., 1985. The duration of the foot-and-mouth disease virus carrier state in african buffalo (i) in the individual animal and (ii) in a free-living herd. *Comp Immunol Microbiol Infect Dis*, **8**(3-4):259–265.

Crandall, K. A., 1995. Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus. *J Virol*, **69**(4):2351–2356.

Crandall, K. A., Kelsey, C. R., Imamichi, H., Lane, H. C., and Salzman, N. P., 1999. Parallel evolution of drug resistance in hiv: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol*, **16**(3):372–382.

Cristina, J. and Colina, R., 2006. Evidence of structural genomic region recombination in hepatitis c virus. *Virol J*, **3**:53.

Crotty, S., Cameron, C. E., and Andino, R., 2001. Rna virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci U S A*, **98**(12):6895–6900.

Cuevas, J. M., Torres-Puente, M., Jimenez-Hernandez, N., Bracho, M. A., Garcia-Robles, I., Wrobel, B., Carnicer, F., del Olmo, J., Ortega, E., Moya, A., *et al.*, 2008. Genetic variability of hepatitis c virus before and after combined therapy of interferon plus ribavirin. *PLoS ONE*, **3**(8):e3058.

Curry, S., Roque-Rosell, N., Sweeney, T. R., Zunszain, P. A., and Leatherbarrow, R. J., 2007. Structural analysis of foot-and-mouth disease virus 3c protease: a viable target for antiviral drugs? *Biochem Soc Trans*, **35**(Pt 3):594–598.

Davidson, F., Simmonds, P., Ferguson, J. C., Jarvis, L. M., Dow, B. C., Follett, E. A., Seed, C. R., Krusius, T., Lin, C., and Medgyesi, G. A., *et al.*, 1995. Survey of major genotypes and subtypes of hepatitis c virus using rflp of sequences amplified from the 5' non-coding region. *J Gen Virol*, **76 (Pt 5)**:1197–1204.

Davis, P. L., Rambaut, A., Bourhy, H., and Holmes, E. C., 2007. The evolutionary dynamics of canid and mongoose rabies virus in southern africa. *Arch Virol*, **152**(7):1251–1258.

de Oliveira, T., Pybus, O. G., Rambaut, A., Salemi, M., Cassol, S., Ciccozzi, M., Rezza, G., Gattinara, G. C., D'Arrigo, R., Amicosante, M., *et al.*, 2006. Molecular epidemiology: Hiv-1 and hcv sequences from libyan outbreak. *Nature*, **444**(7121):836–837.

de Oliveira, T., Salemi, M., Gordon, M., Vandamme, A., van Rensburg, E., Engelbrecht, S., Coovadia, H., and Cassol, S., 2004. Mapping sites of positive selection and amino acid diversification in the hiv genome: an alternative approach to vaccine design? *Genetics*, **167**(3):1047–58–.

DeBry, R. W., Abele, L. G., Weiss, S. H., Hill, M. D., Bouzas, M., Lorenzo, E., Graebnitz, F., and Resnick, L., 1993. Dental hiv transmission? *Nature*, **361**(6414):691.

Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Di Marzio, P., Marmon, S., Sutton, R. E., Hill, C. M., *et al.*, 1996. Identification of a major co-receptor for primary isolates of hiv-1. *Nature*, **381**(6584):661–666.

Deurenberg, R. and Stobberingh, E., 2008. The evolution of staphylococcus aureus. *Infect Genet Evol*, .

Dhanda, M. R., Gopalakrishnan, V. R., and Dhillon, H. S., 1957. Note on the occurrence of atypical strains of foot-and-mouth diseases virus in india. *Ind J Vet Sci*, **27**:79–84.

Dimmic, M. W., Hubisz, M. J., Bustamante, C. D., and Nielsen, R., 2005. Detecting coevolving amino acid sites using bayesian mutational mapping. *Bioinformatics*, **21 Suppl 1**:i126–35.

Dixit, N. M. and Perelson, A. S., 2005. Hiv dynamics with multiple infections of target cells. *Proc Natl Acad Sci U S A*, **102**(23):8198–8203.

Domingo, E., 2000. Viruses at the edge of adaptation. *Virology*, **270**(2):251–253.

Domingo, E., 2002. Quasispecies theory in virology. *J Virol*, **76**(1):463–465.

Domingo, E., Escarmis, C., Lazaro, E., and Manrubia, S., 2005. Quasispecies dynamics and rna virus extinction. *Virus Res*, **107**(2):129–39–.

Domingo, E. and Holland, J. J., 1997. Rna virus mutations and fitness for survival. *Annu Rev Microbiol*, **51**:151–178.

Domingo, E., Martin, V., Perales, C., Grande-Perez, A., Garcia-Arriaza, J., and Arias, A., 2006. Viruses as quasispecies: biological implications. *Curr Top Microbiol Immunol*, **299**:51–82.

Domingo, E., Sabo, D., Taniguchi, T., and Weissmann, C., 1978. Nucleotide sequence heterogeneity of an rna phage population. *Cell*, **13**(4):735–744.

Domingo, E., Verdaguer, N., Ochoa, W., Ruiz-Jarabo, C., Sevilla, N., Baranowski, E., Mateu, M., and Fita, I., 1999. Biochemical and structural studies with neutralizing antibodies raised against foot-and-mouth disease virus. *Virus Res*, **62**(2):169–75–.

Dopazo, J., Rodriguez Minon, J. L., Saiz, J. C., and Sobrino, F., 1995. *Molecular Basis of Virus Evolution*. Cambridge University Press.

Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y., Nagashima, K. A., Cayanan, C., Maddon, P. J., Koup, R. A., Moore, J. P., *et al.*, 1996. Hiv-1 entry into cd4+ cells is mediated by the chemokine receptor cc-ckr-5. *Nature*, **381**(6584):667–673.

Drake, J. W., 1991. A constant rate of spontaneous mutation in dna-based microbes. *Proc Natl Acad Sci U S A*, **88**(16):7160–7164.

Drake, J. W., 1993. Rates of spontaneous mutation among rna viruses. *Proc Natl Acad Sci U S A*, **90**(9):4171–4175.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F., 1998. Rates of spontaneous mutation. *Genetics*, **148**(4):1667–1686.

Drake, J. W. and Holland, J. J., 1999. Mutation rates among rna viruses. *Proc Natl Acad Sci U S A*, **96**(24):13910–13913.

Drake, J. W. and Hwang, C. B. C., 2005. On the mutation rate of herpes simplex virus type 1. *Genetics*, **170**(2):969–970.

Drummond, A., Ho, S., Phillips, M., and Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, **4**(5):e88–.

Drummond, A. and Rambaut, A., 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, **7**(1):214–.

Drummond, A., Rambaut, A., Shapiro, B., and Pybus, O., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, **22**(5):1185–92–.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A., 2003. Measurably evolving populations. *Trends Ecol Evol*, **18**:481–488.

Duffy, S. and Holmes, E. C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded dna begomovirus tomato yellow leaf curl virus. *J Virol*, **82**(2):957–965.

Duffy, S., Shackelton, L. A., and Holmes, E. C., 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*, **9**(4):267–276.

Dutheil, J., Pupko, T., Jean-Marie, A., and Galtier, N., 2005. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol*, **22**(9):1919–1928.

Edgar, R. C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5):1792–1797.

Edwards, C. J., Bollongino, R., Scheu, A., Chamberlain, A., Tresset, A., Vigne, J.-D., Baird, J. F., Larson, G., Ho, S. Y. W., Heupink, T. H., *et al.*, 2007. Mitochondrial dna analysis shows a near eastern neolithic origin for domestic cattle and no indication of domestication of european aurochs. *Proc Biol Sci*, **274**(1616):1377–1385.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, **58**(10):465–523.

Eigen, M., 1987. New concepts for dealing with the evolution of nucleic acids. *Cold Spring Harb Symp Quant Biol*, **52**:307–320.

Eigen, M., 1993. The origin of genetic information: viruses as models. *Gene*, **135**(1-2):37–47.

Eigen, M., 1996a. On the nature of virus quasispecies. *Trends Microbiol*, **4**(6):216–218.

Eigen, M., 1996b. *Steps Towards Life*. Oxford University Press, Oxford.

Eigen, M., 2002. Error catastrophe and antiviral strategy. *Proc Natl Acad Sci U S A*, **99**(21):13374–13376.

Eigen, M. and Schuster, P., 1977. The hypercycle. a principle of natural self-organization. part a: Emergence of the hypercycle. *Naturwissenschaften*, **64**(11):541–565.

Elena, S., Gonzalez-Candelas, F., and Moya, A., 1992. Does the vp1 gene of foot-and-mouth disease virus behave as a molecular clock? *J Mol Evol*, **35**(3):223–9–.

Elena, S. F., Agudelo-Romero, P., Carrasco, P., Codoner, F. M., Martin, S., Torres-Barcelo, C., and Sanjuan, R., 2008. Experimental evolution of plant rna viruses. *Heredity*, **100**(5):478–483.

Elena, S. F., Carrasco, P., Daros, J.-A., and Sanjuan, R., 2006. Mechanisms of genetic robustness in rna viruses. *EMBO Rep*, **7**(2):168–173.

Elena, S. F. and Sanjuan, R., 2005. Adaptive value of high mutation rates of rna viruses: separating causes from consequences. *J Virol*, **79**(18):11555–11558.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., and Beerenwinkel, N., 2008. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, **4**(4):e1000074.

Escarmis, C., Davila, M., Charpentier, N., Bracho, A., Moya, A., and Domingo, E., 1996. Genetic lesions associated with muller's ratchet in an rna virus. *J Mol Biol*, **264**(2):255–267.

Escarmis, C., Lazaro, E., and Manrubia, S. C., 2006. Population bottlenecks in quasispecies dynamics. *Curr Top Microbiol Immunol*, **299**:141–170.

Escriu, F., Fraile, A., and Garcia-Arenal, F., 2007. Constraints to genetic exchange support gene coadaptation in a tripartite rna virus. *PLoS Pathog*, **3**(1):e8.

Farci, P., Alter, H. J., Wong, D. C., Miller, R. H., Govindarajan, S., Engle, R., Shapiro, M., and Purcell, R. H., 1994. Prevention of hepatitis c virus infection in chimpanzees after antibody-mediated in vitro neutralization. *Proc Natl Acad Sci U S A*, **91**(16):7792–7796.

Farci, P., Shimoda, A., Wong, D., Cabezon, T., De Gioannis, D., Strazzera, A., Shimizu, Y., Shapiro, M., Alter, H. J., and Purcell, R. H., *et al.*, 1996. Prevention of hepatitis c virus infection in chimpanzees by hyperimmune serum against the hypervariable region 1 of the envelope 2 protein. *Proc Natl Acad Sci U S A*, **93**(26):15394–15399.

Fares, M., 2004. Swapsc: sliding window analysis procedure to detect selective constraints. *Bioinformatics*, **20**(16):2867–8–.

Fares, M., Elena, S., Ortiz, J., Moya, A., and Barrio, E., 2002a. A sliding window-based method to detect selective constraints in protein-coding genes and its application to rna viruses. *J Mol Evol*, **55**(5):509–21–.

Fares, M. and McNally, D., 2006. Caps: coevolution analysis using protein sequences. *Bioinformatics*, **22**(22):2821–2–.

Fares, M., Moya, A., Escarmis, C., Baranowski, E., Domingo, E., and Barrio, E., 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (fmdv) subjected to experimental passage regimens. *Mol Biol Evol*, **18**(1):10–21–.

Fares, M. and Travers, S., 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, **173**(1):9–23–.

Fares, M. A., 2006. Computational and statistical methods to explore the various dimensions of protein evolution. *Curr Bioinformatics*, **1**:207–217.

Fares, M. A., Ruiz-Gonzalez, M. X., Moya, A., Elena, S. F., and Barrio, E., 2002b. Endosymbiotic bacteria: groel buffers against deleterious mutations. *Nature*, **417**(6887):398.

Fargette, D., Pinel, A., Rakotomalala, M., Sangu, E., Traore, O., Sereme, D., Sorho, F., Issaka, S., Hebrard, E., Sere, Y., *et al.*, 2008. Rice yellow mottle virus, an rna plant virus, evolves as rapidly as most rna animal viruses. *J Virol*, **82**(7):3584–3589.

Fay, J. C. and Wu, C. I., 2001. The neutral theory in the genomic era. *Curr Opin Genet Dev*, **11**(6):642–646.

Fay, J. C. and Wu, C.-I., 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet*, **4**:213–235.

Feng, Y., Broder, C. C., Kennedy, P. E., and Berger, E. A., 1996. Hiv-1 entry cofactor: functional cdna cloning of a seven-transmembrane, g protein-coupled receptor. *Science*, **272**(5263):872–877.

Flanagan, N. S., Tobler, A., Davison, A., Pybus, O. G., Kapan, D. D., Planas, S., Linares, M., Heckel, D., and McMillan, W. O., 2004. Historical demography of mullerian mimicry in the neotropical heliconius butterflies. *Proc Natl Acad Sci U S A*, **101**(26):9704–9709.

Flanegan, J. B. and Baltimore, D., 1977. Poliovirus-specific primer-dependent rna polymerase able to copy poly(a). *Proc Natl Acad Sci U S A*, **74**(9):3677–3680.

Foxman, B. and Riley, L., 2001. Molecular epidemiology: focus on infection. *Am J Epidemiol*, **153**(12):1135–1141.

Fracastoro, G., 1546. *De Contagione et Contagiosis Morbis et eorum curatione.*

Freed, E. O. and Martin, M. A., 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. *J Biol Chem*, **270**(41):23883–23886.

Friedrich, T. C., Frye, C. A., Yant, L. J., O'Connor, D. H., Kriewaldt, N. A., Benson, M., Vojnov, L., Dodds, E. J., Cullen, C., Rudersdorf, R., *et al.*, 2004. Extraepitopic compensatory substitutions partially restore fitness to simian immunodeficiency virus variants that escape from an immunodominant cytotoxic-t-lymphocyte response. *J Virol*, **78**(5):2581–2585.

Fryxell, K. J., 1996. The coevolution of gene family trees. *Trends Genet*, **12**(9):364–369.

Furio, V., Moya, A., and Sanjuan, R., 2005. The cost of replication fidelity in an rna virus. *Proc Natl Acad Sci U S A*, **102**(29):10233–10237.

Furio, V., Moya, A., and Sanjuan, R., 2007. The cost of replication fidelity in human immunodeficiency virus type 1. *Proc Biol Sci*, **274**(1607):225–230.

Galtier, N., 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. *Syst Biol*, **53**(1):38–46.

Gao, F., Bailes, E., Robertson, D., Chen, Y., Rodenburg, C., Michael, S., Cummins, L., Arthur, L., Peeters, M., Shaw, G., *et al.*, 1999. Origin of hiv-1 in the chimpanzee pan troglodytes troglodytes. *Nature*, **397**(6718):436–41–.

Garcia-Arenal, F., Fraile, A., and Malpica, J. M., 2001. Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol*, **39**:157–186.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B., Bhattacharya, T., *et al.*, 2002. Diversity considerations in hiv-1 vaccine selection. *Science*, **296**(5577):2354–60–.

Gaucher, E. A., Gu, X., Miyamoto, M. M., and Benner, S. A., 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci*, **27**(6):315–321.

Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., *et al.*, 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**(7062):1162–1166.

Gibbs, A., Keese, P., Gibbs, M., and Garcia-Arenal, F., 1999. *Origin and evolution of viruses.* Academic Press, San Diego, CA.

Gibbs, A. J., Ohshima, K., Phillips, M. J., and Gibbs, M. J., 2008. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS ONE*, **3**(6):e2523.

Gibbs, M. J., Armstrong, J. S., and Gibbs, A. J., 2000. Sister-scanning: a monte carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, **16**(7):573–582.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K., *et al.*, 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**(5822):222–234.

Gilbert, M. T. P., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M., 2007. The emergence of hiv/aids in the americas and beyond. *Proc Natl Acad Sci U S A*, **104**(47):18566–18570.

Gilbert, P. B., Novitsky, V., and Essex, M., 2005. Covariability of selected amino acid positions for hiv type 1 subtypes c and b. *AIDS Res Hum Retroviruses*, **21**(12):1016–1030.

Gloor, G. B., Martin, L. C., Wahl, L. M., and Dunn, S. D., 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, **44**(19):7156–7165.

Gnanakaran, S., Lang, D., Daniels, M., Bhattacharya, T., Derdeyn, C. A., and Korber, B., 2007. Clade-specific differences between human immunodeficiency virus type 1 clades b and c: diversity and correlations in c3-v4 regions of gp120. *J Virol*, **81**(9):4886–4891.

Gobel, U., Sander, C., Schneider, R., and Valencia, A., 1994. Correlated mutations and residue contacts in proteins. *Proteins*, **18**(4):309–317.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E., 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol*, **299**(2):283–293.

Goldman, N. and Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, **11**(5):725–736.

Gonzalez-Candelas, F., Bracho, M. A., and Moya, A., 2003. Molecular epidemiology and forensic genetics: application to a hepatitis c virus transmission event at a hemodialysis unit. *J Infect Dis*, **187**(3):352–358.

Gorry, P. R., Dunfee, R. L., Mefford, M. E., Kunstman, K., Morgan, T., Moore, J. P., Mascola, J. R., Agopian, K., Holm, G. H., Mehle, A., *et al.*, 2007. Changes in the v3 region of gp120 contribute to unusually broad coreceptor usage of an hiv-1 isolate from a ccr5 delta32 heterozygote. *Virology*, **362**(1):163–178.

Grande-Perez, A., Sierra, S., Castro, M. G., Domingo, E., and Lowenstein, P. R., 2002. Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity. *Proc Natl Acad Sci U S A*, **99**(20):12938–12943.

Grassly, N. C. and Holmes, E. C., 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol*, **14**(3):239–247.

Gray, C. M., Williamson, C., Bredell, H., Puren, A., Xia, X., Filter, R., Zijenah, L., Cao, H., Morris, L., Vardas, E., *et al.*, 2005. Viral dynamics and cd4+ t cell counts in subtype c human immunodeficiency virus type 1-infected individuals from southern africa. *AIDS Res Hum Retroviruses*, **21**(4):285–291.

Grazziotin, F. G., Monzel, M., Echeverrigaray, S., and Bonatto, S. L., 2006. Phylogeography of the bothrops jararaca complex (serpentes: Viperidae): past fragmentation and island colonization in the brazilian atlantic forest. *Mol Ecol*, **15**(13):3969–3982.

Grenfell, B., Pybus, O., Gog, J., Wood, J., Daly, J., Mumford, J., and Holmes, E., 2004a. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**(5656):327–32–.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C., 2004b. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**(5656):327–332.

Griffiths, R. C. and Tavare, S., 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*, **344**(1310):403–410.

Grubman, M. J. and Baxt, B., 2004. Foot-and-mouth disease. *Clin Microbiol Rev*, **17**(2):465–493.

Gu, X., 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol*, **16**(12):1664–1674.

Gu, X., 2001. Mathematical modeling for functional divergence after gene duplication. *J Comput Biol*, **8**(3):221–234.

Guillot, S., Caro, V., Cuervo, N., Korotkova, E., Combiescu, M., Persu, A., Aubert-Combiescu, A., Delpeyroux, F., and Crainic, R., 2000. Natural genetic exchanges between vaccine and wild poliovirus strains in humans. *J Virol*, **74**(18):8434–8443.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P., 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*, **101**(35):12957–12962.

Gutierrez, M. I., Spangler, G., Kingma, D., Raffeld, M., Guerrero, I., Misad, O., Jaffe, E. S., Magrath, I. T., and Bhatia, K., 1998. Epstein-barr virus in nasal lymphomas contains multiple ongoing mutations in the ebna-1 gene. *Blood*, **92**(2):600–606.

Haas, J., Park, E. C., and Seed, B., 1996. Codon usage limitation in the expression of hiv-1 envelope glycoprotein. *Curr Biol*, **6**(3):315–324.

Hall, A., 1996. What is molecular epidemiology? *Trop Med Int Health*, **1**(4):407–408.

Hanada, K., Suzuki, Y., and Gojobori, T., 2004. A large variation in the rates of synonymous substitution for rna viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol*, **21**(6):1074–1080.

Hartl, F. U., 1996. Molecular chaperones in cellular protein folding. *Nature*, **381**(6583):571–579.

Hartl, F. U., Hlodan, R., and Langer, T., 1994. Molecular chaperones in protein folding: the art of avoiding sticky situations. *Trends Biochem Sci*, **19**(1):20–25.

Hartley, O., Klasse, P. J., Sattentau, Q. J., and Moore, J. P., 2005. V3: Hiv's switch-hitter. *AIDS Res Hum Retroviruses*, **21**(2):171–189.

Hatta, M., Halfmann, P., Wells, K., and Kawaoka, Y., 2002. Human influenza a viral genes responsible for the restriction of its replication in duck intestine. *Virology*, **295**(2):250–255.

Haydon, D., Bastos, A., Knowles, N., and Samuel, A., 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics*, **157**(1):7–15–.

Hayward, A. and Stone, G. N., 2006. Comparative phylogeography across two trophic levels: the oak gall wasp andricus kollari and its chalcid parasitoid megastigmus stigmatizans. *Mol Ecol*, **15**(2):479–489.

He, C.-Q., Ding, N.-Z., Chen, J.-G., and Li, Y.-L., 2007. Evidence of natural recombination in classical swine fever virus. *Virus Res*, **126**(1-2):179–185.

Heath, L., van der Walt, E., Varsani, A., and Martin, D. P., 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol*, **80**(23):11827–11832.

Henderson, W., 1978. An histroical review of the control of foot-and-mouth diseas. *Br vet J*, **134**:3–9.

Henikoff, S. and Henikoff, J. G., 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**(22):10915–10919.

Hill, W. G. and Robertson, A., 1966. The effect of linkage on limits to artificial selection. *Genet Res*, **8**(3):269–294.

Hillis, D. M. and Huelsenbeck, J. P., 1994. Support for dental hiv transmission. *Nature*, **369**(6475):24–25.

Hmaied, F., Legrand-Abravanel, F., Nicot, F., Garrigues, N., Chapuy-Regaud, S., Dubois, M., Njouom, R., Izopet, J., and Pasquier, C., 2007. Full-length genome sequences of hepatitis c virus subtype 4f. *J Gen Virol*, **88**(Pt 11):2985–2990.

Hogle, J. M., 2002. Poliovirus cell entry: common structural themes in viral cell entry pathways. *Annu Rev Microbiol*, **56**:677–702.

Holder, M. and Lewis, P. O., 2003. Phylogeny estimation: traditional and bayesian approaches. *Nat Rev Genet*, **4**(4):275–284.

Holland, J. J., Domingo, E., de la Torre, J. C., and Steinhauer, D. A., 1990. Mutation frequencies at defined single codon sites in vesicular stomatitis virus and poliovirus can be increased only slightly by chemical mutagenesis. *J Virol*, **64**(8):3960–3962.

Hollier, M. J. and Dimmock, N. J., 2005. The c-terminal tail of the gp41 transmembrane envelope glycoprotein of hiv-1 clades a, b, c, and d may exist in two conformations: an analysis of sequence, structure, and function. *Virology*, **337**(2):284–296.

Hollister, J. R., Vagnozzi, A., Knowles, N. J., and Rieder, E., 2008. Molecular and phylogenetic analyses of bovine rhinovirus type 2 shows it is closely related to foot-and-mouth disease virus. *Virology*, **373**(2):411–425.

Holmes, E., 1998a. The molecular archaeology of infectious disease. *SGM Quarterly*, :56–57.

Holmes, E., 2007. When hiv spread afar. *Proc Natl Acad Sci U S A*, :–.

Holmes, E. C., 1998b. Molecular epidemiology and evolution of emerging infectious diseases. *Br Med Bull*, **54**(3):533–543.

Holmes, E. C., 2003. Error thresholds and the constraints to rna virus evolution. *Trends Microbiol*, **11**(12):543–546.

Holmes, E. C., 2004. The phylogeography of human viruses. *Mol Ecol*, **13**(4):745–756.

Holmes, E. C., 2008. Evolutionary history and phylogeography of human viruses. *Annu Rev Microbiol*, **62**:307–328.

Holmes, E. C., Brown, A. J., and Simmonds, P., 1993a. Sequence data as evidence. *Nature*, **364**(6440):766.

Holmes, E. C. and Drummond, A. J., 2007. The evolutionary genetics of viral emergence. *Curr Top Microbiol Immunol*, **315**:51–66.

Holmes, E. C. and Moya, A., 2002. Is the quasispecies concept relevant to rna viruses? *J Virol*, **76**(1):460–465.

Holmes, E. C. and Twiddy, S. S., 2003. The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol*, **3**(1):19–28.

Holmes, E. C., Worobey, M., and Rambaut, A., 1999. Phylogenetic evidence for recombination in dengue virus. *Mol Biol Evol*, **16**(3):405–409.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A., and Brown, A. J., 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc Natl Acad Sci U S A*, **89**(11):4835–4839.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Rogers, A. S., and Brown, A. J., 1993b. Molecular investigation of human immunodeficiency virus (hiv) infection in a patient of an hiv-infected surgeon. *J Infect Dis*, **167**(6):1411–1414.

Huang, C.-c., Tang, M., Zhang, M.-Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I. A., *et al.*, 2005. Structure of a v3-containing hiv-1 gp120 core. *Science*, **310**(5750):1025–1028.

Hudson, R. R., 1983a. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, **23**(2):183–201.

Hudson, R. R., 1983b. Testing the constant rate neutral allele model with protein sequence data. *Evolution*, **37**:203–217.

Hue, S., Pillay, D., Clewley, J., and Pybus, O., 2005. Genetic analysis reveals the complex structure of hiv-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*, **102**(12):4425–9–.

Hughes, G., 2008. A reassessment of the emergence time of european bat lyssavirus type 1. *Infect Genet Evol*, .

Hwang, S. S., Boyle, T. J., Lyerly, H. K., and Cullen, B. R., 1991. Identification of the envelope v3 loop as the primary determinant of cell tropism in hiv-1. *Science*, **253**(5015):71–74.

Ina, Y., 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol*, **40**(2):190–226.

Inoue, T., Parida, S., Paton, D. J., Linchongsubongkoch, W., Mackay, D., Oh, Y., Aunpomma, D., Gubbins, S., and Saeki, T., 2006. Development and evaluation of an indirect enzyme-linked immunosorbent assay for detection of foot-and-mouth disease virus nonstructural protein antibody using a chemically synthesized 2b peptide as antigen. *J Vet Diagn Invest*, **18**(6):545–552.

Jackson, A., O'Neill, H., Maree, F., Blignaut, B., Carrillo, C., Rodriguez, L., and Haydon, D., 2007. Mosaic structure of foot-and-mouth disease virus genomes. *J Gen Virol*, **88**(Pt 2):487–92–.

Jackson, T., Ellard, F. M., Ghazaleh, R. A., Brookes, S. M., Blakemore, W. E., Corteyn, A. H., Stuart, D. I., Newman, J. W., and King, A. M., 1996. Efficient infection of cells in culture by type o foot-and-mouth disease virus requires binding to cell surface heparan sulfate. *J Virol*, **70**(8):5282–5287.

Jackson, T., Mould, A. P., Sheppard, D., and King, A. M. Q., 2002. Integrin alphavbeta1 is a receptor for foot-and-mouth disease virus. *J Virol*, **76**(3):935–941.

Jackson, T., Sheppard, D., Denyer, M., Blakemore, W., and King, A. M., 2000. The epithelial integrin alphavbeta6 is a receptor for foot-and-mouth disease virus. *J Virol*, **74**(11):4949–4956.

Jenkins, G. M., Rambaut, A., Pybus, O. G., and Holmes, E. C., 2002. Rates of molecular evolution in rna viruses: a quantitative phylogenetic analysis. *J Mol Evol*, **54**(2):156–165.

Jenkins, G. M., Worobey, M., Woelk, C. H., and Holmes, E. C., 2001. Evidence for the non-quasispecies evolution of rna viruses [corrected]. *Mol Biol Evol*, **18**(6):987–994.

Jockusch, H., Wiegand, C., Mersch, B., and Rajes, D., 2001. Mutants of tobacco mosaic virus with temperature-sensitive coat proteins induce heat shock response in tobacco leaves. *Mol Plant Microbe Interact*, **14**(7):914–917.

Jones, L. R. and Weber, E. L., 2004. Homologous recombination in bovine pestiviruses. phylogenetic and statistic evidence. *Infect Genet Evol*, **4**(4):335–343.

Jung, A., Maier, R., Vartanian, J.-P., Bocharov, G., Jung, V., Fischer, U., Meese, E., Wain-Hobson, S., and Meyerhans, A., 2002. Multiply infected spleen cells in hiv patients. *Nature*, **418**(6894):144.

Kadam, U. and Ghosh, S., 2008. Codon adaptation index analysis of rna genome plant viruses. *Current Science*, **94**(1):24–27.

Kageyama, S., Agdamag, D. M., Alesna, E. T., Leano, P. S., Heredia, A. M. L., Abellanosa-Tac-An, I. P., Jereza, L. D., Tanimoto, T., Yamamura, J.-i., and Ichimura, H., *et al.*, 2006. A natural inter-genotypic (2b/1b) recombinant of hepatitis c virus in the philippines. *J Med Virol*, **78**(11):1423–1428.

Kalinina, O., Norder, H., Mukomolov, S., and Magnius, L. O., 2002. A natural intergenotypic recombinant of hepatitis c virus identified in st. petersburg. *J Virol*, **76**(8):4034–4043.

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., *et al.*, 2006. Chimpanzee reservoirs of pandemic and nonpandemic hiv-1. *Science*, **313**(5786):523–526.

Kew, O., Morris-Glasgow, V., Landaverde, M., Burns, C., Shaw, J., Garib, Z., Andre, J., Blackman, E., Freeman, C. J., Jorba, J., *et al.*, 2002. Outbreak of poliomyelitis in hispaniola associated with circulating type 1 vaccine-derived poliovirus. *Science*, **296**(5566):356–359.

Kijak, G. H. and McCutchan, F. E., 2005. Hiv diversity, molecular epidemiology, and the role of recombination. *Curr Infect Dis Rep*, **7**(6):480–488.

Kilbourne, E. D., 1973. The molecular epidemiology of influenza. *J Infect Dis*, **127**(4):478–487.

Kim, Y. and Subramaniam, S., 2006. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins*, **62**(4):1115–1124.

Kimura, M., 1968. Evolutionary rate at the molecular level. *Nature*, **217**(5129):624–626.

Kimura, M., 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, **267**(5608):275–276.

King, J. L. and Jukes, T. H., 1969. Non-darwinian evolution. *Science*, **164**(881):788–798.

Kingman, J. F., 1982. On the geneaology of large populations. *J Appl Prob*, **19a**:27–43.

Kingman, J. F., 2000. Origins of the coalescent. 1974-1982. *Genetics*, **156**(4):1461–1463.

Knowles, N. and Samuel, A., 2003. Molecular epidemiology of foot-and-mouth disease virus. *Virus Res*, **91**(1):65–80–.

Koelle, K., Cobey, S., Grenfell, B., and Pascual, M., 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science*, **314**(5807):1898–1903.

Kondrashov, A. S., 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature*, **336**(6198):435–440.

Konig, G., Blanco, C., Knowles, N., Palma, E., Maradei, E., and Piccone, M., 2001. Phylogenetic analysis of foot-and-mouth disease viruses isolated in argentina. *Virus Genes*, **23**(2):175–81–.

Korber, B., Farber, R., Wolpert, D., and Lapedes, A., 1993. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A*, **90**(15):7176–80–.

Korber, B., MacInnes, K., Smith, R., and Myers, G., 1994. Mutational trends in v3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. *J Virol*, **68**(10):6730–44–.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B., Wolinsky, S., and Bhattacharya, T., 2000. Timing the ancestor of the hiv-1 pandemic strains. *Science*, **288**(5472):1789–96–.

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. W., 2006a. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol*, **23**(10):1891–1901.

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. W., 2006b. Gard: a genetic algorithm for recombination detection. *Bioinformatics*, **22**(24):3096–3098.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A., 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet*, **4**(8):e1000144.

Kreitman, M., 2000. Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet*, **1**:539–559.

Kuhner, M. K., Yamato, J., and Felsenstein, J., 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**(1):429–434.

Kumar, S., Tamura, K., and Nei, M., 2004. Mega3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*, **5**(2):150–163.

Kuntzen, T., Berical, A., Ndjomou, J., Bennett, P., Schneidewind, A., Lennon, N., Birren, B. W., Kuiken, C., Henn, M. R., Simmonds, P., *et al.*, 2008. A set of reference sequences for the hepatitis c genotypes 4d, 4f, and 4k covering the full open reading frame. *J Med Virol*, **80**(8):1370–1378.

Kurbanov, F., Tanaka, Y., Kramvis, A., Simmonds, P., and Mizokami, M., 2008. When should "i" consider a new hepatitis b virus genotype? *J Virol*, **82**(16):8241–8242.

Kwong, P. D., Wyatt, R., Majeed, S., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A., 2000. Structures of hiv-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure*, **8**(12):1329–1339.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A., 1998. Structure of an hiv gp120 envelope glycoprotein in complex with the cd4 receptor and a neutralizing human antibody. *Nature*, **393**(6686):648–659.

Lai, M. M., 1992. Rna recombination in animal and plant viruses. *Microbiol Rev*, **56**(1):61–79.

Le Vaillant, F., 1795. *Second Voyage dans l'Interieur de l'Afrique, par le Cap de Bonne-Esperance, dans les annees 1783, 84 et 85.* Paris: H.J.Jansen.

Lea, S., Hernandez, J., Blakemore, W., Brocchi, E., Curry, S., Domingo, E., Fry, E., Abu-Ghazaleh, R., King, A., Newman, J., *et al.*, 1994. The structure and antigenicity of a type c foot-and-mouth disease virus. *Structure*, **2**(2):123–39–.

Lee, C. H., Gilbertson, D. L., Novella, I. S., Huerta, R., Domingo, E., and Holland, J. J., 1997. Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus. *J Virol*, **71**(5):3636–3640.

Lefeuvre, P., Lett, J.-M., Reynaud, B., and Martin, D. P., 2007. Avoidance of protein fold disruption in natural virus recombinants. *PLoS Pathog*, **3**(11):e181.

Legrand-Abravanel, F., Claudinon, J., Nicot, F., Dubois, M., Chapuy-Regaud, S., Sandres-Saune, K., Pasquier, C., and Izopet, J., 2007. New natural intergenotypic (2/5) recombinant of hepatitis c virus. *J Virol*, **81**(8):4357–4362.

Lemey, P., Pybus, O., Wang, B., Saksena, N., Salemi, M., and Vandamme, A., 2003. Tracing the origin and history of the hiv-2 epidemic. *Proc Natl Acad Sci U S A*, **100**(11):6588–92–.

Lemey, P., Rambaut, A., and Pybus, O. G., 2006. Hiv evolutionary dynamics within and among hosts. *AIDS Rev*, **8**(3):125–140.

Levi, J. E., Takaoka, D. T., Garrini, R. H., Fachini, R. M., Focaccia, R., De Bortholi Santos, E., Mitre, H. P., De Mendonca, J. S., De Paula Cavalheiro, N., Barone, A. A., *et al.*, 2002. Three cases of infection with hepatitis c virus genotype 5 among brazilian hepatitis patients. *J Clin Microbiol*, **40**(7):2645–2647.

Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Leigh Brown, A. J., 2008. Episodic sexual transmission of hiv revealed by molecular phylodynamics. *PLoS Med*, **5**(3):e50.

Lewis-Rogers, N., McClellan, D., and Crandall, K., 2008. The evolution of foot-and-mouth disease virus: Impacts of recombination and selection. *Infect Genet Evol*, .

Li, W. H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*, **36**(1):96–99.

Li, W. H., Wu, C. I., and Luo, C. C., 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, **2**(2):150–174.

Li, Y., Carroll, D., Gardner, S., Walsh, M., Vitalis, E., and Damon, I., 2007. On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc Natl Acad Sci U S A*, **104**(40):15787–92–.

Liu, W., Worobey, M., Li, Y., Keele, B. F., Bibollet-Ruche, F., Guo, Y., Goepfert, P. A., Santiago, M. L., Ndjango, J.-B. N., Neel, C., *et al.*, 2008. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. *PLoS Pathog*, **4**(7):e1000097.

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. *Mol Biol Evol*, **15**(9):1183–1188.

Loeb, L. A., Essigmann, J. M., Kazazi, F., Zhang, J., Rose, K. D., and Mullins, J. I., 1999. Lethal mutagenesis of hiv with mutagenic nucleoside analogs. *Proc Natl Acad Sci U S A*, **96**(4):1492–1497.

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W., and Ray, S. C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype c-infected seroconverters in india, with evidence of intersubtype recombination. *J Virol*, **73**(1):152–160.

Lopez-Labrador, F. X., Moya, A., and Gonzalez-Candelas, F., 2008. Mapping natural polymorphisms of hepatitis c virus ns3/4a protease and antiviral resistance to inhibitors in worldwide isolates. *Antivir Ther*, **13**(4):481–494.

Lu, L., Li, C., Fu, Y., Gao, F., Pybus, O. G., Abe, K., Okamoto, H., Hagedorn, C. H., and Murphy, D., 2007a. Complete genomes of hepatitis c virus (hcv) subtypes 6c, 6l, 6o, 6p and 6q: completion of a full panel of genomes for hcv genotype 6. *J Gen Virol*, **88**(Pt 5):1519–1525.

Lu, L., Li, C., Fu, Y., Thaikruea, L., Thongswat, S., Maneekarn, N., Apichartpiyakul, C., Hotta, H., Okamoto, H., Netski, D., *et al.*, 2007b. Complete genomes for hepatitis c virus subtypes 6f, 6i, 6j and 6m: viral genetic diversity among thai blood donors and infected spouses. *J Gen Virol*, **88**(Pt 5):1505–1518.

Lu, L., Li, C., and Hagedorn, C. H., 2006a. Phylogenetic analysis of global hepatitis e virus sequences: genetic diversity, subtypes and zoonosis. *Rev Med Virol*, **16**(1):5–36.

Lu, L., Murphy, D., Li, C., Liu, S., Xia, X., Pham, P. H., Jin, Y., Hagedorn, C. H., and Abe, K., 2008. Complete genomes of three subtype 6t isolates and analysis of many novel hepatitis c virus variants within genotype 6. *J Gen Virol*, **89**(Pt 2):444–452.

Lu, L., Nakano, T., Li, C., Fu, Y., Miller, S., Kuiken, C., Robertson, B. H., and Hagedorn, C. H., 2006b. Hepatitis c virus complete genome sequences identified from china representing subtypes 6k and 6n and a novel, as yet unassigned subtype within genotype 6. *J Gen Virol*, **87**(Pt 3):629–634.

Lukashov, V. V. and Goudsmit, J., 2002. Recent evolutionary history of human immunodeficiency virus type 1 subtype b: reconstruction of epidemic onset based on sequence distances to the common ancestor. *J Mol Evol*, **54**(5):680–691.

Maddison, W. and Maddison, D., 1992. Macclade. *Sunderland (MA): Sinauer Associates*, .

Magiorkinis, G., Paraskevis, D., Magiorkinis, E., Vandamme, A.-M., and Hatzakis, A., 2005. Re-analysis of 34 full-length hiv-1 intersubtype recombinant sequences. *Infect Genet Evol*, **5**(3):225–229.

Mahy, B. W. J., 2005. Introduction and history of foot-and-mouth disease virus. *Curr Top Microbiol Immunol*, **288**:1–8.

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W., and Garcia-Arenal, F., 2002. The rate and character of spontaneous mutation in an rna virus. *Genetics*, **162**(4):1505–1511.

Mansky, L. M., 2000. In vivo analysis of human t-cell leukemia virus type 1 reverse transcription accuracy. *J Virol*, **74**(20):9525–9531.

Mansky, L. M. and Bernard, L. C., 2000. 3'-azido-3'-deoxythymidine (azt) and azt-resistant reverse transcriptase can increase the in vivo mutation rate of human immunodeficiency virus type 1. *J Virol*, **74**(20):9532–9539.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., *et al.*, 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057):376–380.

Martin, D. and Rybicki, E., 2000. Rdp: detection of recombination amongst aligned sequences. *Bioinformatics*, **16**(6):562–563.

Martin, D. P., Posada, D., Crandall, K. A., and Williamson, C., 2005a. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses*, **21**(1):98–102.

Martin, D. P., van der Walt, E., Posada, D., and Rybicki, E. P., 2005b. The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet*, **1**(4):e51.

Martin, D. P., Williamson, C., and Posada, D., 2005c. Rdp2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**(2):260–262.

Martinez, M. A., Dopazo, J., Hernandez, J., Mateu, M. G., Sobrino, F., Domingo, E., and Knowles, N. J., 1992. Evolution of the capsid protein genes of foot-and-mouth disease virus: antigenic variation without accumulation of amino acid substitutions over six decades. *J Virol*, **66**(6):3557–3565.

Martinez, M. A., Verdaguer, N., Mateu, M. G., and Domingo, E., 1997. Evolution subverting essentiality: dispensability of the cell attachment arg-gly-asp motif in multiply passaged foot-and-mouth disease virus. *Proc Natl Acad Sci U S A*, **94**(13):6798–6802.

Martinez-Solano, I., Jockusch, E. L., and Wake, D. B., 2007. Extreme population subdivision throughout a continuous range: phylogeography of batrachoseps attenuatus (caudata: Plethodontidae) in western north america. *Mol Ecol*, **16**(20):4335–4355.

Mason, P., Pacheco, J., Zhao, Q., and Knowles, N., 2003. Comparisons of the complete genomes of asian, african and european isolates of a recent foot-and-mouth disease virus type o pandemic strain (panasia). *J Gen Virol*, **84**(Pt 6):1583–93–.

Mateu, M. G., Valero, M. L., Andreu, D., and Domingo, E., 1996. Systematic replacement of amino acid residues within an arg-gly-asp-containing loop of foot-and-mouth disease virus and effect on cell recognition. *J Biol Chem*, **271**(22):12814–12819.

Mayer, M. P., 2005. Recruitment of hsp70 chaperones: a crucial part of viral survival strategies. *Rev Physiol Biochem Pharmacol*, **153**:1–46.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T., 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, **23**(13):i319–27.

McCormack, G. P. and Clewley, J. P., 2002. The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Rev Med Virol*, **12**(4):221–238.

McNally, D. and Fares, M. A., 2007. In silico identification of functional divergence between the multiple groel gene paralogs in chlamydiae. *BMC Evol Biol*, **7**:81.

Menendez-Arias, L., Martinez, M. A., Quinones-Mateu, M. E., and Martinez-Picado, J., 2003. Fitness variations and their impact on the evolution of antiretroviral drug resistance. *Curr Drug Targets Infect Disord*, **3**(4):355–371.

Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptak, R. G., Gibbs, R. A., and Hillis, D. M., 2002. Molecular evidence of hiv-1 transmission in a criminal case. *Proc Natl Acad Sci U S A*, **99**(22):14292–14297.

Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z.-G., and Arnold, F. H., 2003. Library analysis of schema-guided protein recombination. *Protein Sci*, **12**(8):1686–1693.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A., 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol*, **25**(7):1459–1471.

Mittal, M., Tosh, C., Hemadri, D., Sanyal, A., and Bandyopadhyay, S. K., 2005. Phylogeny, genome evolution, and antigenic variability among endemic foot-and-mouth disease virus type a isolates from india. *Arch Virol*, **150**(5):911–928.

Miyata, T. and Yasunaga, T., 1980. Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, **16**(1):23–36.

Moffat, K., Howell, G., Knox, C., Belsham, G., Monaghan, P., Ryan, M., and Wileman, T., 2005. Effects of foot-and-mouth disease virus nonstructural proteins on the structure and function of the early secretory pathway: 2bc but not 3a blocks endoplasmic reticulum-to-golgi transport. *J Virol*, **79**(7):4382–95–.

Montville, R., Froissart, R., Remold, S. K., Tenaillon, O., and Turner, P. E., 2005. Evolution of mutational robustness in an rna virus. *PLoS Biol*, **3**(11):e381.

Moreno, I. M., Malpica, J. M., Diaz-Pendon, J. A., Moriones, E., Fraile, A., and Garcia-Arenal, F., 2004. Variability and genetic structure of the population of watermelon mosaic virus infecting melon in spain. *Virology*, **318**(1):451–460.

Moriyama, E. N. and Powell, J. R., 1997. Synonymous substitution rates in drosophila: mitochondrial versus nuclear genes. *J Mol Evol*, **45**(4):378–391.

Moya, A., Holmes, E. C., and Gonzalez-Candelas, F., 2004. The population genetics and evolutionary epidemiology of rna viruses. *Nat Rev Microbiol*, **2**(4):279–288.

Muller, H. J., 1964. The relation of recombination to mutational advance. *Mutat Res*, **106**:2–9.

Murphy, D., Chamberland, J., Dandavino, R., et al., 2007a. A new genotype of hepatitis c virus originating from central africa. In *58th Annual Meeting of the American Association for the Study of Liver Diseases (AASLD)*, number 872.

Murphy, D., Willems, B., and Delage, G., 1994. Use of the 5' noncoding region for genotyping hepatitis c virus. *J Infect Dis*, **169**(2):473–475.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R., and Sabbah, S., 2007b. Use of sequence analysis of the ns5b region for routine genotyping of hepatitis c virus with reference to c/e1 and 5' untranslated region sequences. *J Clin Microbiol*, **45**(4):1102–1112.

Murray, C. L., Jones, C. T., and Rice, C. M., 2008. Architects of assembly: roles of flaviviridae non-structural proteins in virion morphogenesis. *Nat Rev Microbiol*, **6**(9):699–708.

Muse, S. V. and Gaut, B. S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, **11**(5):715–724.

Nakano, T., Lu, L., He, Y., Fu, Y., Robertson, B. H., and Pybus, O. G., 2006. Population genetic history of hepatitis c virus 1b infection in china. *J Gen Virol*, **87**(Pt 1):73–82.

Ndung'u, T., Lu, Y., Renjifo, B., Touzjian, N., Kushner, N., Pena-Cruz, V., Novitsky, V. A., Lee, T. H., and Essex, M., 2001. Infectious simian/human immunodeficiency virus with human immunodeficiency virus type 1 subtype c from an african isolate: rhesus macaque model. *J Virol*, **75**(23):11417–11425.

Nei, M. and Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, **3**(5):418–426.

Neilson, J. R., John, G. C., Carr, J. K., Lewis, P., Kreiss, J. K., Jackson, S., Nduati, R. W., Mbori-Ngacha, D., Panteleeff, D. D., Bodrug, S., *et al.*, 1999. Subtypes of human immunodeficiency virus type 1 and disease stage among women in nairobi, kenya. *J Virol*, **73**(5):4397–4403.

Nelson, M. I., Edelman, L., Spiro, D. J., Boyne, A. R., Bera, J., Halpin, R., Ghedin, E., Miller, M. A., Simonsen, L., Viboud, C., *et al.*, 2008. Molecular epidemiology of a/h3n2 and a/h1n1 influenza virus during a single epidemic season in the united states. *PLoS Pathog*, **4**(8):e1000133.

Nielsen, R., 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**(Pt 6):641–647.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., *et al.*, 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, **3**(6):e170.

Nielsen, R. and Weinreich, D. M., 1999. The age of nonsynonymous and synonymous mutations in animal mtdna and implications for the mildly deleterious theory. *Genetics*, **153**(1):497–506.

Nielsen, R. and Yang, Z., 1998a. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, **148**(3):929–36–.

Nielsen, R. and Yang, Z., 1998b. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, **148**(3):929–936.

Nisole, S., Stoye, J. P., and Saib, A., 2005. Trim family proteins: retroviral restriction and antiviral defence. *Nat Rev Microbiol*, **3**(10):799–808.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessiere, A., Ayouba, A., Pasquier, C., and Pouillot, R., 2006. The hepatitis c virus epidemic in cameroon: Genetic evidence for rapid transmission between 1920 and 1960. *Infect Genet Evol*, :–.

Noppornpanth, S., Lien, T. X., Poovorawan, Y., Smits, S. L., Osterhaus, A. D. M. E., and Haagmans, B. L., 2006. Identification of a naturally occurring recombinant genotype 2/6 hepatitis c virus. *J Virol*, **80**(15):7569–7577.

Noppornpanth, S., Poovorawan, Y., Lien, T. X., Smits, S. L., Osterhaus, A. D. M. E., and Haagmans, B. L., 2008. Complete genome analysis of hepatitis c virus subtypes 6t and 6u. *J Gen Virol*, **89**(Pt 5):1276–1281.

Notredame, C., Higgins, D. G., and Heringa, J., 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1):205–217.

Novella, I. S., Duarte, E. A., Elena, S. F., Moya, A., Domingo, E., and Holland, J. J., 1995. Exponential increases of rna virus fitness during large population transmissions. *Proc Natl Acad Sci U S A*, **92**(13):5841–5844.

Nowak, M. A., 1992. What is a quasispecies? *Trends Ecol Evol*, **7**(4):118–121.

Nunez, J., Baranowski, E., Molina, N., Ruiz-Jarabo, C., Sanchez, C., Domingo, E., and Sobrino, F., 2001. A single amino acid substitution in nonstructural protein 3a can mediate adaptation of foot-and-mouth disease virus to the guinea pig. *J Virol*, **75**(8):3977–83–.

Ochoa, W. F., Kalko, S. G., Mateu, M. G., Gomes, P., Andreu, D., Domingo, E., Fita, I., and Verdaguer, N., 2000. A multiply substituted g-h loop from foot-and-mouth disease virus in complex with a neutralizing antibody: a role for water molecules. *J Gen Virol*, **81**(Pt 6):1495–1505.

Ohno, H., Aguilar, R. C., Fournier, M. C., Hennecke, S., Cosson, P., and Bonifacino, J. S., 1997. Interaction of endocytic signals from the hiv-1 envelope glycoprotein complex with members of the adaptor medium chain family. *Virology*, **238**(2):305–315.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, **246**(5428):96–98.

Ohta, T., 1992. Theoretical study of near neutrality. ii. effect of subdivided population structure with local extinction and recolonization. *Genetics*, **130**(4):917–923.

Ohta, T., 2002. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A*, **99**(25):16134–16137.

Otto, C., Puffer, B. A., Pohlmann, S., Doms, R. W., and Kirchhoff, F., 2003. Mutations in the c3 region of human and simian immunodeficiency virus envelope have differential effects on viral infectivity, replication, and cd4-dependency. *Virology*, **315**(2):292–302.

Otto, S. P. and Lenormand, T., 2002. Resolving the paradox of sex and recombination. *Nat Rev Genet*, **3**(4):252–261.

Ou, C. Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C. C., Korber, B. T., Mullins, J. I., Schochetman, G., Berkelman, R. L., and Economou, A. N., *et al.*, 1992. Molecular epidemiology of hiv transmission in a dental practice. *Science*, **256**(5060):1165–1171.

Pacheco, J. M., Henry, T. M., O'Donnell, V. K., Gregory, J. B., and Mason, P. W., 2003. Role of nonstructural proteins 3a and 3b in host range and pathogenicity of foot-and-mouth disease virus. *J Virol*, **77**(24):13017–13027.

Padidam, M., Sawyer, S., and Fauquet, C. M., 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology*, **265**(2):218–225.

Pamilo, P. and Bianchi, N. O., 1993. Evolution of the zfx and zfy genes: rates and interdependence between the genes. *Mol Biol Evol*, **10**(2):271–281.

Pantophlet, R. and Burton, D. R., 2006. Gp120: target for neutralizing hiv-1 antibodies. *Annu Rev Immunol*, **24**:739–769.

Pape, J. W., Farmer, P., Koenig, S., Fitzgerald, D., Wright, P., and Johnson, W., 2008. The epidemiology of aids in haiti refutes the claims of gilbert et al. *Proc Natl Acad Sci U S A*, **105**(10):E13.

Pardi, F. and Goldman, N., 2005. Species choice for comparative genomics: being greedy works. *PLoS Genet*, **1**(6):e71.

Parmley, J. L., Chamary, J. V., and Hurst, L. D., 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol*, **23**(2):301–309.

Parry, N., Fox, G., Rowlands, D., Brown, F., Fry, E., Acharya, R., Logan, D., and Stuart, D., 1990. Structural and serological evidence for a novel mechanism of antigenic variation in foot-and-mouth disease virus. *Nature*, **347**(6293):569–72–.

Pazos, F., Ranea, J. A. G., Juan, D., and Sternberg, M. J. E., 2005. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol*, **352**(4):1002–1015.

Pedersen, A. K., Wiuf, C., and Christiansen, F. B., 1998. A codon-based model designed to describe lentiviral evolution. *Mol Biol Evol*, **15**(8):1069–1081.

Peeters, M. and Sharp, P. M., 2000. Genetic diversity of hiv-1: the moving target. *AIDS*, **14 Suppl 3**:S129–40.

Pei, J., Cai, W., Kinch, L. N., and Grishin, N. V., 2006. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**(2):164–171.

Pereda, A. J., Konig, G. A., Chimeno Zoth, S. A., Borca, M., Palma, E. L., and Piccone, M. E., 2002. Full length nucleotide sequence of foot-and-mouth disease virus strain o1 campos/bra/58. brief report. *Arch Virol*, **147**(11):2225–2230.

Pereira, H., 1991. *Foot and Mouth Disease*. Academic Press Inc, London, United Kingdom.

Perez-Losada, M., Porter, M. L., Tazi, L., and Crandall, K. A., 2007. New methods for inferring population dynamics from microbial sequences. *Infect Genet Evol*, **7**(1):24–43.

Perez-Losada, M., Viscidi, R. P., Demma, J. C., Zenilman, J., and Crandall, K. A., 2005. Population genetics of neisseria gonorrhoeae in a high-prevalence community using a hypervariable outer membrane porb and 13 slowly evolving housekeeping genes. *Mol Biol Evol*, **22**(9):1887–1902.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J., 1980. The evolution of genes: the chicken preproinsulin gene. *Cell*, **20**(2):555–566.

Perno, C. F., Svicher, V., and Ceccherini-Silberstein, F., 2006. Novel drug resistance mutations in hiv: recognition and clinical relevance. *AIDS Rev*, **8**(4):179–190.

Pfeiffer, J. K. and Kirkegaard, K., 2003. A single mutation in poliovirus rna-dependent rna polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. *Proc Natl Acad Sci U S A*, **100**(12):7289–7294.

Pfeiffer, J. K. and Kirkegaard, K., 2005. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog*, **1**(2):e11.

Piana, S., Carloni, P., and Rothlisberger, U., 2002. Drug resistance in hiv-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci*, **11**(10):2393–2402.

Pillay, D., Rambaut, A., Geretti, A. M., and Brown, A. J. L., 2007. Hiv phylogenetics. *BMJ*, **335**(7618):460–461.

Ping, L. H., Nelson, J. A., Hoffman, I. F., Schock, J., Lamers, S. L., Goodman, M., Vernazza, P., Kazembe, P., Maida, M., Zimba, D., *et al.*, 1999. Characterization of v3 sequence heterogeneity in subtype c human immunodeficiency virus type 1 isolates from malawi: underrepresentation of x4 variants. *J Virol*, **73**(8):6271–6281.

Poignard, P., Saphire, E. O., Parren, P. W., and Burton, D. R., 2001. gp120: Biologic aspects of structural features. *Annu Rev Immunol*, **19**:253–274.

Pollock, D. D. and Taylor, W. R., 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng*, **10**(6):647–657.

Pollock, D. D., Taylor, W. R., and Goldman, N., 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, **287**(1):187–198.

Pond, S. L. K. and Frost, S. D. W., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, **21**(10):2531–2533.

Poon, A., Lewis, F., Pond, S., and Frost, S., 2007. Evolutionary interactions between n-linked glycosylation sites in the hiv-1 envelope. *PLoS Comput Biol*, **3**(1):e11–.

Porter, A. G., 1993. Picornavirus nonstructural proteins: emerging roles in virus replication and inhibition of host cell functions. *J Virol*, **67**(12):6917–6921.

Posada, D., 2000. How does recombination affect phylogeny estimation? *Trends Ecol Evol*, **15**(12):489–490.

Posada, D., 2002. Evaluation of methods for detecting recombination from dna sequences: empirical data. *Mol Biol Evol*, **19**(5):708–717.

Posada, D. and Crandall, K. A., 1998. Modeltest: testing the model of dna substitution. *Bioinformatics*, **14**(9):817–818.

Posada, D. and Crandall, K. A., 2001. Evaluation of methods for detecting recombination from dna sequences: computer simulations. *Proc Natl Acad Sci U S A*, **98**(24):13757–13762.

Posada, D. and Crandall, K. A., 2002. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol*, **54**(3):396–402.

Posada, D., Crandall, K. A., and Holmes, E. C., 2002. Recombination in evolutionary genomics. *Annu Rev Genet*, **36**:75–97.

Pouillot, R., Lachenal, G., Pybus, O., Rousset, D., and Njouom, R., 2008. Variable epidemic histories of hepatitis c virus genotype 2 infection in west africa and cameroon. *Infect Genet Evol*, .

Pritchard, L., Bladon, P., M O Mitchell, J., and J Dufton, M., 2001. Evaluation of a novel method for the identification of coevolving protein residues. *Protein Eng*, **14**(8):549–555.

Pugachev, K. V., Guirakhoo, F., Ocran, S. W., Mitchell, F., Parsons, M., Penal, C., Girakhoo, S., Pougatcheva, S. O., Arroyo, J., Trent, D. W., *et al.*, 2004. High fidelity of yellow fever virus rna polymerase. *J Virol*, **78**(2):1032–1038.

Pybus, O., Drummond, A., Nakano, T., Robertson, B., and Rambaut, A., 2003a. The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: a bayesian coalescent approach. *Mol Biol Evol*, **20**(3):381–7–.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., and Harvey, P. H., 2001. The epidemic behavior of the hepatitis c virus. *Science*, **292**(5525):2323–2325.

Pybus, O. G., Cochrane, A., Holmes, E. C., and Simmonds, P., 2005. The hepatitis c virus epidemic among injecting drug users. *Infect Genet Evol*, **5**(2):131–139.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H., and Rambaut, A., 2003b. The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: a bayesian coalescent approach. *Mol Biol Evol*, **20**(3):381–387.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J., and Holmes, E. C., 2007. Phylogenetic evidence for deleterious mutation load in rna viruses and its contribution to viral evolution. *Mol Biol Evol*, **24**(3):845–852.

Pybus, O. G., Rambaut, A., and Harvey, P. H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**(3):1429–1437.

Ramani, A. K. and Marcotte, E. M., 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, **327**(1):273–284.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C., 2008. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, **453**(7195):615–619.

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J., and Thomas, D. L., 2000. Genetic epidemiology of hepatitis c virus throughout egypt. *J Infect Dis*, **182**(3):698–707.

Reich, D. E. and Goldstein, D. B., 1998. Genetic evidence for a paleolithic human population expansion in africa. *Proc Natl Acad Sci U S A*, **95**(14):8119–8123.

Renjifo, B., Gilbert, P., Chaplin, B., Vannberg, F., Mwakagile, D., Msamanga, G., Hunter, D., Fawzi, W., and Essex, M., 1999. Emerging recombinant human immunodeficiency viruses: uneven representation of the envelope v3 region. *Aids*, **13**(13):1613–21–.

Resch, A. M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A. Y., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V., 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol*, **24**(8):1821–1831.

Resch, W., Hoffman, N., and Swanstrom, R., 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, **288**(1):51–62.

Rizzuto, C. D., Wyatt, R., Hernandez-Ramos, N., Sun, Y., Kwong, P. D., Hendrickson, W. A., and Sodroski, J., 1998. A conserved hiv gp120 glycoprotein structure involved in chemokine receptor binding. *Science*, **280**(5371):1949–1953.

Robbins, K. E., Lemey, P., Pybus, O. G., Jaffe, H. W., Youngpairoj, A. S., Brown, T. M., Salemi, M., Vandamme, A.-M., and Kalish, M. L., 2003. U.s. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J Virol*, **77**(11):6359–6366.

Robertson, B., Myers, G., Howard, C., Brettin, T., Bukh, J., Gaschen, B., Gojobori, T., Maertens, G., Mizokami, M., Nainan, O., *et al.*, 1998. Classification, nomenclature, and database development for hepatitis c virus (hcv) and related viruses: proposals for standardization. international committee on virus taxonomy. *Arch Virol*, **143**(12):2493–2503.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., *et al.*, 2000. Hiv-1 nomenclature proposal. *Science*, **288**(5463):55–56.

Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H., 1995. Recombination in hiv-1. *Nature*, **374**(6518):124–126.

Rogaev, E. I., Moliaka, Y. K., Malyarchuk, B. A., Kondrashov, F. A., Derenko, M. V., Chumakov, I., and Grigorenko, A. P., 2006. Complete mitochondrial genome and phylogeny of pleistocene mammoth mammuthus primigenius. *PLoS Biol*, **4**(3):e73.

Roingeard, P., Hourioux, C., Blanchard, E., Brand, D., and Ait-Goughoulte, M., 2004. Hepatitis c virus ultrastructure and morphogenesis. *Biol Cell*, **96**(2):103–108.

Rosenberg, N. A. and Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, **3**(5):380–390.

Ross, R. S., Viazov, S., Renzing-Kohler, K., and Roggendorf, M., 2000. Changes in the epidemiology of hepatitis c infection in germany: shift in the predominance of hepatitis c subtypes. *J Med Virol*, **60**(2):122–125.

Roth, C., Betts, M. J., Steffansson, P., Saelensminde, G., and Liberles, D. A., 2005. The adaptive evolution database (taed): a phylogeny based tool for comparative genomics. *Nucleic Acids Res*, **33**(Database issue):D495–7.

Rowell, J. F., Stanhope, P. E., and Siliciano, R. F., 1995. Endocytosis of endogenously synthesized hiv-1 envelope protein. mechanism and role in processing for association with class ii mhc. *J Immunol*, **155**(1):473–488.

Ruano-Rubio, V. and Fares, M. A., 2007. Testing the neutral fixation of hetero-oligomerism in the archaeal chaperonin cct. *Mol Biol Evol*, **24**(6):1384–1396.

Rueckert, R., 1996. *Picornaviridae: the viruses and their replication.* Lippincott-Raven, Philadelphia, PA.

Ruiz-Jarabo, C., Sevilla, N., Davila, M., Gomez-Mariano, G., Baranowski, E., and Domingo, E., 1999. Antigenic properties and population stability of a foot-and-mouth disease virus with an altered arg-gly-asp receptor-recognition motif. *J Gen Virol*, **80 (Pt 8)**:1899–909–.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**(4):406–425.

Salemi, M., Lewis, M., Egan, J. F., Hall, W. W., Desmyter, J., and Vandamme, A. M., 1999. Different population dynamics of human t cell lymphotropic virus type ii in intravenous drug users compared with endemically infected tribes. *Proc Natl Acad Sci U S A*, **96**(23):13253–13258.

Salemi, M. and Vandamme, A.-M., 2002. Hepatitis c virus evolutionary patterns studied through analysis of full-genome sequences. *J Mol Evol*, **54**(1):62–70.

Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E., 1995. Identification of breakpoints in intergenotypic recombinants of hiv type 1 by bootscanning. *AIDS Res Hum Retroviruses*, **11**(11):1423–1425.

Sangare, O., Bastos, A., Marquardt, O., Venter, E., Vosloo, W., and Thomson, G., 2001. Molecular epidemiology of serotype o foot-and-mouth disease virus with emphasis on west and south africa. *Virus Genes*, **22**(3):345–51–.

Sangare, O., Bastos, A., Venter, E., and Vosloo, W., 2003. Retrospective genetic analysis of sat-1 type foot-and-mouth disease outbreaks in west africa (1975-1981). *Vet Microbiol*, **93**(4):279–89–.

Sanjuan, R., Cuevas, J. M., Furio, V., Holmes, E. C., and Moya, A., 2007. Selection for robustness in mutagenized rna viruses. *PLoS Genet*, **3**(6):e93.

Sanjuan, R. and Elena, S. F., 2006. Epistasis correlates to genomic complexity. *Proc Natl Acad Sci U S A*, **103**(39):14402–14405.

Sanjuan, R., Forment, J., and Elena, S. F., 2006. In silico predicted robustness of viroid rna secondary structures. ii. interaction between mutation pairs. *Mol Biol Evol*, **23**(11):2123–2130.

Sanjuan, R., Moya, A., and Elena, S. F., 2004a. The contribution of epistasis to the architecture of fitness in an rna virus. *Proc Natl Acad Sci U S A*, **101**(43):15376–15379.

Sanjuan, R., Moya, A., and Elena, S. F., 2004b. The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proc Natl Acad Sci U S A*, **101**(22):8396–8401.

Santiago, M., Range, F., Keele, B., Li, Y., Bailes, E., Bibollet-Ruche, F., Fruteau, C., Noe, R., Peeters, M., Brookfield, J., *et al.*, 2005. Simian immunodeficiency virus infection in free-ranging sooty mangabeys (cercocebus atys atys) from the tai forest, cote d'ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *J Virol*, **79**(19):12515–27–.

Sardanyes, J., Elena, S. F., and Sole, R. V., 2008. Simple quasispecies models for the survival-of-the-flattest effect: The role of space. *J Theor Biol*, **250**(3):560–568.

Schierup, M. H. and Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**(2):879–891.

Seibert, S., Howell, C., Hughes, M., and Hughes, A., 1995. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (hiv-1). *Mol Biol Evol*, **12**(5):803–13–.

Sentandreu, V., Jimenez-Hernandez, N., Torres-Puente, M., Bracho, M. A., Valero, A., Gosalbes, M. J., Ortega, E., Moya, A., and Gonzalez-Candelas, F., 2008. Evidence of recombination in intrapatient populations of hepatitis c virus. *PLoS ONE*, **3**(9):e3239.

Severson, W. E., Schmaljohn, C. S., Javadian, A., and Jonsson, C. B., 2003. Ribavirin causes error catastrophe during hantaan virus replication. *J Virol*, **77**(1):481–488.

Sewatanon, J., Srichatrapimuk, S., and Auewarakul, P., 2007. Compositional bias and size of genomes of human dna viruses. *Intervirology*, **50**(2):123–132.

Shackelton, L. A. and Holmes, E. C., 2006. Phylogenetic evidence for the rapid evolution of human b19 erythrovirus. *J Virol*, **80**(7):3666–3669.

Shackelton, L. A., Parrish, C. R., Truyen, U., and Holmes, E. C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci U S A*, **102**(2):379–384.

Shapiro, B., Drummond, A., Rambaut, A., Wilson, M., Matheus, P., Sher, A., Pybus, O., Gilbert, M., Barnes, I., Binladen, J., *et al.*, 2004. Rise and fall of the beringian steppe bison. *Science*, **306**(5701):1561–5–.

Shapiro, B., Rambaut, A., Pybus, O. G., and Holmes, E. C., 2006. A phylogenetic method for detecting positive epistasis in gene sequences and its application to rna virus evolution. *Mol Biol Evol*, **23**(9):1724–1730.

Sharp, P., Bailes, E., Chaudhuri, R., Rodenburg, C., Santiago, M., and Hahn, B., 2001. The origins of acquired immune deficiency syndrome viruses: where and when? *Philos Trans R Soc Lond B Biol Sci*, **356**(1410):867–76–.

Sharp, P. M., 1997. In search of molecular darwinism. *Nature*, **385**(6612):111–112.

Sharp, P. M. and Li, W. H., 1987. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**(3):1281–1295.

Sharp, P. M., Stenico, M., Peden, J. F., and Lloyd, A. T., 1993. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans*, **21**(4):835–841.

Sheridan, I., Pybus, O. G., Holmes, E. C., and Klenerman, P., 2004. High-resolution phylogenetic analysis of hepatitis c virus adaptation and its relationship to disease progression. *J Virol*, **78**(7):3447–3454.

Shimizu, Y. K., Hijikata, M., Iwamoto, A., Alter, H. J., Purcell, R. H., and Yoshikura, H., 1994. Neutralizing antibodies against hepatitis c virus and the emergence of neutralization escape mutant viruses. *J Virol*, **68**(3):1494–1500.

Shindyalov, I. N., Kolchanov, N. A., and Sander, C., 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*, **7**(3):349–358.

Shriner, D., Nickle, D. C., Jensen, M. A., and Mullins, J. I., 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res*, **81**(2):115–121.

Shriner, D., Rodrigo, A. G., Nickle, D. C., and Mullins, J. I., 2004. Pervasive genomic recombination of hiv-1 in vivo. *Genetics*, **167**(4):1573–1583.

Sierra, M., Airaksinen, A., Gonzalez-Lopez, C., Agudo, R., Arias, A., and Domingo, E., 2007. Foot-and-mouth disease virus mutant with decreased sensitivity to ribavirin: implications for error catastrophe. *J Virol*, **81**(4):2012–2024.

Sierra, S., Davila, M., Lowenstein, P. R., and Domingo, E., 2000. Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity. *J Virol*, **74**(18):8316–8323.

Simmonds, P., 2004. Genetic diversity and evolution of hepatitis c virus–15 years on. *J Gen Virol*, **85**(Pt 11):3173–3188.

Simmonds, P., 2006. Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded rna viruses. *J Virol*, **80**(22):11124–11140.

Simmonds, P., Bukh, J., Combet, C., Deleage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspe, G., Kuiken, C., Maertens, G., *et al.*, 2005. Consensus proposals for a unified system of nomenclature of hepatitis c virus genotypes. *Hepatology*, **42**(4):962–973.

Simmonds, P., Holmes, E. C., Cha, T. A., Chan, S. W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J., and Urdea, M. S., *et al.*, 1993. Classification of hepatitis c virus into six major genotypes and a series of subtypes by phylogenetic analysis of the ns-5 region. *J Gen Virol*, **74 (Pt 11)**:2391–2399.

Simmonds, P. and Midgley, S., 2005. Recombination in the genesis and evolution of hepatitis b virus genotypes. *J Virol*, **79**(24):15467–15476.

Simmons, H. E., Holmes, E. C., and Stephenson, A. G., 2008. Rapid evolutionary dynamics of zucchini yellow mosaic virus. *J Gen Virol*, **89**(Pt 4):1081–1085.

Smith, J. M., 1992. Analyzing the mosaic structure of genes. *J Mol Evol*, **34**(2):126–129.

Smith, T. F. and Waterman, M. S., 1992. The continuing case of the florida dentist. *Science*, **256**(5060):1155–1156.

Sobrino, F., Saiz, M., Jimenez-Clavero, M., Nunez, J., Rosas, M., Baranowski, E., and Ley, V., 2001. Foot-and-mouth disease virus: a long known virus, but a current threat. *Vet Res*, **32**(1):1–30–.

Steinhauer, D. A., Domingo, E., and Holland, J. J., 1992. Lack of evidence for proofreading mechanisms associated with an rna virus polymerase. *Gene*, **122**(2):281–288.

Steinhauer, D. A. and Holland, J. J., 1987. Rapid evolution of rna viruses. *Annu Rev Microbiol*, **41**:409–433.

Stephens, M., 2001. *Handbook of Statistical Genetics*. John Wiley and Sons, UK.

Stevens, R. D., Robinson, A. J., and Goble, C. A., 2003. mygrid: personalised bioinformatics on the information grid. *Bioinformatics*, **19 Suppl 1**:i302–4.

Stevens, R. D., Tipney, H. J., Wroe, C. J., Oinn, T. M., Senger, M., Lord, P. W., Goble, C. A., Brass, A., and Tassabehji, M., 2004. Exploring williams-beuren syndrome using mygrid. *Bioinformatics*, **20 Suppl 1**:i303–10.

Storz, J. F., Beaumont, M. A., and Alberts, S. C., 2002. Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical bayesian model. *Mol Biol Evol*, **19**(11):1981–1990.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S., 2001. Bayesian selection of continuous-time markov chain evolutionary models. *Mol Biol Evol*, **18**(6):1001–1013.

Suzuki, Y., 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol Biol Evol*, **21**(12):2352–2359.

Suzuki, Y., Gojobori, T., and Nei, M., 2001. Adaptsite: detecting natural selection at single amino acid sites. *Bioinformatics*, **17**(7):660–661.

Suzuki, Y. and Nei, M., 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol Biol Evol*, **18**(12):2179–2185.

Sweeney, T. R., Roque-Rosell, N., Birtley, J. R., Leatherbarrow, R. J., and Curry, S., 2007. Structural and mutagenic analysis of foot-and-mouth disease virus 3c protease reveals the role of the beta-ribbon in proteolysis. *J Virol*, **81**(1):115–124.

Switzer, W., Salemi, M., Shanmugam, V., Gao, F., Cong, M., Kuiken, C., Bhullar, V., Beer, B., Vallet, D., Gautier-Hion, A., *et al.*, 2005. Ancient co-speciation of simian foamy viruses and primates. *Nature*, **434**(7031):376–80–.

Swofford, D., 1998. Paup*: phylogenetic analysis using parsimony (* and other methods). *Sunderland (MA): Sinauer Associates*, .

Taboga, O., Tami, C., Carrillo, E., Nunez, J. I., Rodriguez, A., Saiz, J. C., Blanco, E., Valero, M. L., Roig, X., Camarero, J. A., *et al.*, 1997. A large-scale evaluation of peptide vaccines against foot-and-mouth disease: lack of solid protection in cattle and isolation of escape mutants. *J Virol*, **71**(4):2606–2614.

Tajima, F., 1983. Evolutionary relationship of dna sequences in finite populations. *Genetics*, **105**(2):437–460.

Taubenberger, J. K., Reid, A. H., Lourens, R. M., Wang, R., Jin, G., and Fanning, T. G., 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature*, **437**(7060):889–893.

Taylor, W. R. and Hatrick, K., 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng*, **7**(3):341–348.

Tillier, E. and Collins, R., 1995. Neighbor joining and maximum likelihood with rna sequences: Addressing the interdependence of sites. *Mol Biol Evol*, **12**(1):7–15.

Tillier, E. R. M. and Lui, T. W. H., 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**(6):750–755.

Timm, J., Neukamm, M., Kuntzen, T., Kim, A. Y., Chung, R. T., Brander, C., Lauer, G. M., Walker, B. D., and Allen, T. M., 2007. Characterization of full-length hepatitis c virus genotype 4 sequences. *J Viral Hepat*, **14**(5):330–337.

Torres-Puente, M., Cuevas, J. M., Jimenez-Hernandez, N., Bracho, M. A., Garcia-Robles, I., Carnicer, F., del Olmo, J., Ortega, E., Moya, A., and Gonzalez-Candelas, F., *et al.*, 2008a. Hepatitis c virus and the controversial role of the interferon sensitivity determining region in the response to interferon treatment. *J Med Virol*, **80**(2):247–253.

Torres-Puente, M., Cuevas, J. M., Jimenez-Hernandez, N., Bracho, M. A., Garcia-Robles, I., Wrobel, B., Carnicer, F., Del Olmo, J., Ortega, E., Moya, A., *et al.*, 2008b. Genetic variability in hepatitis c virus and its role in antiviral treatment response. *J Viral Hepat*, **15**(3):188–199.

Tosh, C., Hemadri, D., Sanyal, A., and Bandyopadhyay, S., 2003. Genetic and antigenic analysis of two recently circulating genotypes of type a foot-and-mouth disease virus in india: evidence for positive selection in the capsid-coding genes. *Arch Virol*, **148**(5):853–69–.

Tosh, C., Sanyal, A., Hemadri, D., and Venkataramanan, R., 2002. Phylogenetic analysis of serotype a foot-and-mouth disease virus isolated in india between 1977 and 2000. *Arch Virol*, **147**(3):493–513–.

Travers, S., O'Connell, M., McCormack, G., and McInerney, J., 2005. Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes. *J Virol*, **79**(3):1836–41–.

Travers, S. A. A. and Fares, M. A., 2007. Functional coevolutionary networks of the hsp70-hop-hsp90 system revealed through computational analyses. *Mol Biol Evol*, **24**(4):1032–1044.

Trkola, A., Purtscher, M., Muster, T., Ballaun, C., Buchacher, A., Sullivan, N., Srinivasan, K., Sodroski, J., Moore, J. P., and Katinger, H., *et al.*, 1996. Human monoclonal antibody 2g12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *J Virol*, **70**(2):1100–1108.

Tuffley, C. and Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci*, **147**(1):63–91.

Tully, D. C. and Fares, M. A., 2006. Unravelling selection shifts among foot-and-mouth disease virus serotypes. *Evolutionary Bioinformatics*, **2**:223–237.

Twiddy, S. S. and Holmes, E. C., 2003. The extent of homologous recombination in members of the genus flavivirus. *J Gen Virol*, **84**(Pt 2):429–440.

UNAIDS and WHO, 2007. Aids epidemic update. Technical report.

Vakharia, V. N., Devaney, M. A., Moore, D. M., Dunn, J. J., and Grubman, M. J., 1987. Proteolytic processing of foot-and-mouth disease virus polyproteins expressed in a cell-free system from clone-derived transcripts. *J Virol*, **61**(10):3199–3207.

Vallee, H. and Carre, H., 1922. Sur la pluralite des virus aphtexu. *CR Acad Sci Paris*, **174**:1498–1500.

Vallender, E. J. and Lahn, B. T., 2004. Positive selection on the human genome. *Hum Mol Genet*, **13 Spec No 2**:R245–54.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., *et al.*, 2006. Human immunodeficiency viruses: Siv infection in wild gorillas. *Nature*, **444**(7116):164–.

van Rensburg, H., Haydon, D., Joubert, F., Bastos, A., Heath, L., and Nel, L., 2002. Genetic heterogeneity in the foot-and-mouth disease virus leader and 3c proteinases. *Gene*, **289**(1-2):19–29–.

Vandamme, A. M., Bertazzoni, U., and Salemi, M., 2000. Evolutionary strategies of human t-cell lymphotropic virus type ii. *Gene*, **261**(1):171–180.

Vasilakis, N., Holmes, E., Fokam, E., Faye, O., Diallo, M., Sall, A., and Weaver, S., 2007. Evolutionary processes among sylvatic dengue-2 viruses. *J Virol*, :–.

Verbeeck, J., Maes, P., Lemey, P., Pybus, O. G., Wollants, E., Song, E., Nevens, F., Fevery, J., Delport, W., Van der Merwe, S., *et al.*, 2006. Investigating the origin and spread of hepatitis c virus genotype 5a. *J Virol*, **80**(9):4220–4226.

Verbeeck, J., Peigue-Lafeuille, H., Ross, R. S., Abergel, A., Nevens, F., Van der Merwe, S., Van Ranst, M., and Henquell, C., 2008. Hcv genotype 5: epidemiology and spread of an uncommon genotype. *J Clin Virol*, **41**(2):170–171.

Verdaguer, N., Mateu, M. G., Andreu, D., Giralt, E., Domingo, E., and Fita, I., 1995. Structure of the major antigenic loop of foot-and-mouth disease virus complexed with a neutralizing antibody: direct involvement of the arg-gly-asp motif in the interaction. *EMBO J*, **14**(8):1690–1696.

Verdaguer, N., Schoehn, G., Ochoa, W. F., Fita, I., Brookes, S., King, A., Domingo, E., Mateu, M. G., Stuart, D., and Hewat, E. A., *et al.*, 1999. Flexibility of the major antigenic loop of foot-and-mouth disease virus bound to a fab fragment of a neutralising antibody: structure and neutralisation. *Virology*, **255**(2):260–268.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., and Delaporte, E., *et al.*, 2000. Unprecedented degree of human immunodeficiency virus type 1 (hiv-1) group m genetic diversity in the democratic republic of congo suggests that the hiv-1 pandemic originated in central africa. *J Virol*, **74**(22):10498–507–.

Vignuzzi, M., Stone, J. K., and Andino, R., 2005. Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications. *Virus Res*, **107**(2):173–181.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R., 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, **439**(7074):344–348.

Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H., 2002. Protein building blocks preserved by recombination. *Nat Struct Biol*, **9**(7):553–558.

Vosloo, W., Bastos, A., Kirkbride, E., Esterhuysen, J., van Rensburg, D., Bengis, R., Keet, D., and Thomson, G., 1996. Persistent infection of african buffalo (syncerus caffer) with sat-type foot-and-mouth disease viruses: rate of fixation of mutations, antigenic change and interspecies transmission. *J Gen Virol*, **77 (Pt 7)**:1457–67–.

Wain, L., Bailes, E., Bibollet-Ruche, F., Decker, J., Keele, B., Van Heuverswyn, F., Li, Y., Takehisa, J., Mpoudi Ngole, E., Shaw, G., *et al.*, 2007. Adaptation of hiv-1 to its human host. *Mol Biol Evol*, :–.

Waldmann, O. and Trautwein, K., 1926. Experimentelle untersuchungen über die pluralität des maul-und klauenseuchevirus. *Berl Tierärztl Wschr*, **42**:569–571.

Walker, P., Pybus, O., Rambaut, A., and Holmes, E., 2005. Comparative population dynamics of hiv-1 subtypes b and c: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol*, **5**(3):199–208–.

Wall, J. D., 2000. A comparison of estimators of the population recombination rate. *Mol Biol Evol*, **17**(1):156–163.

Walsh, P. D., Biek, R., and Real, L. A., 2005. Wave-like spread of ebola zaire. *PLoS Biol*, **3**(11):e371.

Wang, Y. and Gu, X., 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**(3):1311–1320.

Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., Salazar-Gonzalez, J. F., Salazar, M. G., Kilby, J. M., Saag, M. S., *et al.*, 2003. Antibody neutralization and escape by hiv-1. *Nature*, **422**(6929):307–312.

Weiner, A. J., Geysen, H. M., Christopherson, C., Hall, J. E., Mason, T. J., Saracco, G., Bonino, F., Crawford, K., Marion, C. D., and Crawford, K. A., *et al.*, 1992. Evidence for immune selection of hepatitis c virus (hcv) putative envelope glycoprotein variants: potential role in chronic hcv infections. *Proc Natl Acad Sci U S A*, **89**(8):3468–3472.

Weinstock, J., Willerslev, E., Sher, A., Tong, W., Ho, S. Y. W., Rubenstein, D., Storer, J., Burns, J., Martin, L., Bravi, C., *et al.*, 2005. Evolution, systematics, and phylogeography of pleistocene horses in the new world: a molecular perspective. *PLoS Biol*, **3**(8):e241.

Westfall, P. and Young, S., 1993. Resampling-based multiple testing. *John Wiley and Sons, New York*, .

Whelan, S., Lio, P., and Goldman, N., 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, **17**(5):262–272.

WHO, 1999. Hepatitis c–global prevalence (update). *Wkly Epidemiol Rec*, **74**(49):425–427.

WHO, 2004. The world health report 2004 - changing history.

Wilke, C. O., 2001. Selection for fitness versus selection for robustness in rna secondary structure folding. *Evolution*, **55**(12):2412–2420.

Wilke, C. O., 2005. Quasispecies theory in the context of population genetics. *BMC Evol Biol*, **5**:44.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C., 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, **412**(6844):331–333.

Wilson, I. J., Weale, M. E., and Balding, D. J., 2003. Inferences from dna data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A*, **166**:155–188.

Wiuf, C., Christensen, T., and Hein, J., 2001. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*, **18**(10):1929–1939.

Woolhouse, M. E. J., 2002. Population biology of emerging and re-emerging pathogens. *Trends Microbiol*, **10**(10 Suppl):S3–7.

Worobey, M., 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol Biol Evol*, **18**(8):1425–1434.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., *et al.*, 2008a. Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, **455**(7213):661–664.

Worobey, M. and Holmes, E. C., 1999. Evolutionary aspects of recombination in rna viruses. *J Gen Virol*, **80 (Pt 10)**:2535–2543.

Worobey, M., Pitchenik, A. E., Gilbert, M. T. P., Wlasiuk, G., and Rambaut, A., 2008b. Reply to pape et al.: the phylogeography of hiv-1 group m subtype b. *Proc Natl Acad Sci U S A*, **105**(12):E16.

Worobey, M., Rambaut, A., and Holmes, E. C., 1999. Widespread intra-serotype recombination in natural populations of dengue virus. *Proc Natl Acad Sci U S A*, **96**(13):7352–7357.

Wyatt, R., Desjardin, E., Olshevsky, U., Nixon, C., Binley, J., Olshevsky, V., and Sodroski, J., 1997. Analysis of the interaction of the human immunodeficiency virus type 1 gp120 envelope glycoprotein with the gp41 transmembrane glycoprotein. *J Virol*, **71**(12):9722–9731.

Wyatt, R., Kwong, P. D., Desjardins, E., Sweet, R. W., Robinson, J., Hendrickson, W. A., and Sodroski, J. G., 1998. The antigenic structure of the hiv gp120 envelope glycoprotein. *Nature*, **393**(6686):705–711.

Yagnik, A. T., Lahm, A., Meola, A., Roccasecca, R. M., Ercole, B. B., Nicosia, A., and Tramontano, A., 2000. A model for the hepatitis c virus envelope glycoprotein e2. *Proteins*, **40**(3):355–366.

Yamaguchi-Kabata, Y. and Gojobori, T., 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J Virol*, **74**(9):4335–50–.

Yang, W., Bielawski, J., and Yang, Z., 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol*, **57**(2):212–21–.

Yang, Z., 1997. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, **13**(5):555–556.

Yang, Z., 2001. Maximum likelihood analysis of adaptive evolution in hiv-1 gp120 env gene. *Pac Symp Biocomput*, :226–37–.

Yang, Z., 2006. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution.

Yang, Z. and Bielawski, J., 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, **15**(12):496–503.

Yang, Z. and Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**(1):32–43.

Yang, Z. and Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, **19**(6):908–917.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1):431–49–.

Yang, Z., Wong, W. S. W., and Nielsen, R., 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, **22**(4):1107–1118.

Yeh, W. W., Cale, E. M., Jaru-Ampornpan, P., Lord, C. I., Peyerl, F. W., and Letvin, N. L., 2006. Compensatory substitutions restore normal core assembly in simian immunodeficiency virus isolates with gag epitope cytotoxic t-lymphocyte escape mutations. *J Virol*, **80**(16):8168–8177.

Yokoyama, S., 2002. Molecular evolution of color vision in vertebrates. *Gene*, **300**(1-2):69–78.

Zhang, C., Ding, N., and Wei, J., 2008. Different sliding window sizes and inappropriate subtype references result in discordant mosaic maps and breakpoint locations of hiv-1 crfs. *Infect Genet Evol*, .

Zhang, J., 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*, **21**(7):1332–1339.

Zhang, J., Nielsen, R., and Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**(12):2472–2479.

Zheng, Y., Roberts, R. J., and Kasif, S., 2002. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol*, **3**(11):RESEARCH0060.

Zhou, S., Liu, R., Baroudy, B. M., Malcolm, B. A., and Reyes, G. R., 2003. The effect of ribavirin and impdh inhibitors on hepatitis c virus subgenomic replicon rna. *Virology*, **310**(2):333–342.

Zhou, W., Parent, L., Wills, J., and Resh, M., 1994. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 gag protein which interacts with acidic phospholipids. *J Virol*, **68**(4):2556–69–.

Zhou, Y. and Holmes, E., 2007. Bayesian estimates of the evolutionary rate and age of hepatitis b virus. *J Mol Evol*, :–.

Zibert, A., Kraas, W., Meisel, H., Jung, G., and Roggendorf, M., 1997. Epitope mapping of antibodies directed against hypervariable region 1 in acute self-limiting and chronic infections due to hepatitis c virus. *J Virol*, **71**(5):4123–4127.