Within-host evolution of HIV-1 and the analysis of transmissible diversity

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

By

Suzanne English

Wolfson College and

Nuffield Department of Clinical Medicine

University of Oxford

Hilary Term 2012



Contents

CONTENTS **2**

FIGURES **8**

TABLES **15**

ABSTRACT **19**

STATEMENT OF AUTHORSHIP **20**

ACKNOWLEDGMENTS **21**

CHAPTER 1 INTRODUCTION **22**

1.1 OVERVIEW 22

1.1.1 Central problems in HIV-1 vaccine research *22*

1.1.2 The apparent design of HIV-1 for transmission *23*

1.1.3 Non adaptive variation, sampling error and bias *24*

1.1.4 The effective population size *25*

1.1.5 Hard selection at transmission *26*

1.1.6 Acute infection and infectiousness *27*

1.1.7 Problems with inferring single variant infection *27*

1.1.8 Alternatives to single variant transmission *28*

1.2 THEORETICAL CONCEPTS UNDERPINNING THE EXPERIMENTAL APPROACH 31

1.2.1 Dynamical models *31*

1.2.2 The effect of diploidy on HIV-1 evolution *31*

1.2.3 Quantifying natural selection *32*

1.2.4 Stochastic versus adaptive changes *35*

1.2.5 Interpreting natural selection *36*

1.2.6 Quantifying relatedness by common descent *38*

1.3 TRANSMISSION 40

1.3.1 Overview *40*

1.3.2 The biology of human transmission *41*

1.3.3 Non-human models used to infer key events in HIV-1 transmission *50*

1.3.4 Risk factors for human-to-human transmission and viral evolution *59*

1.3.5 Defining adaptive qualities for transmission *61*

1.3.6 Transmission bottlenecks *62*

1.4 HIV-1 DISEASE AND EVOLUTION 64
1.4.1 Overview *64*
1.4.2 Pathogenicity of HIV-1 ..... *64*
1.4.3 Host susceptibility *67*
1.5 MULTIPLE-INFECTION 71
1.5.1 Overview and definitions *71*
1.5.2 Prevalence of multiple-infection *72*
1.5.3 Factors which limit multiple-infection of a host *72*
1.6 HOMOLOGOUS RECOMBINATION 74
1.6.1 Overview *74*
1.6.2 Mechanisms of homologous recombination *74*
1.6.3 Recombination and the evolutionary origins of HIV-1 ..... *77*
1.6.4 Consideration of some theoretical disadvantages of recombination for HIV-1 adaptation within hosts *80*
1.7 SUMMARY OF WORK 82
1.7.1 Hypothesis *82*
1.7.2 Identification of key questions *82*
1.7.3 An outline of the contents of each chapter *83*
1.7.4 Summary of experimental data *84*
CHAPTER 2 MATERIALS AND METHODS 86
2.1 OVERVIEW 86
2.2 SUBJECTS 87
2.2.1 Definitions and abbreviations *87*
2.2.2 Cohorts *87*
2.2.3 Fiebig Staging *88*
2.2.4 Ethics Statement *92*
2.3 LABORATORY METHODS 93
2.3.1 Definitions and abbreviations *93*
2.3.2 Sampling of sequences within hosts *93*
2.3.3 Steps taken to prevent contamination *95*
2.4 ANALYTICAL METHODS 96
2.4.1 Abbreviations *96*
2.4.2 Sequence alignment *96*
2.4.3 Phylogenetic analysis *98*

2.4.4 Time of most recent common ancestor (tMRCA) estimation *108*

2.4.5 Recombination analysis *112*

CHAPTER 3 INFERRING THE CONTEMPORANEOUS TRANSMISSION OF DIVERGENT VIRAL VARIANTS FROM A SINGLE DONOR TO TWO RECIPIENTS ..**114**

3.1 OVERVIEW 114

3.1.1 Brief summary of chapter and hypothesis *114*

3.1.2 Key abbreviations *116*

3.2 INTRODUCTION 117

3.2.1 Antigenicity and vaccine design *117*

3.2.2 Genetic bottleneck *117*

3.2.3 Possible confounding factors *118*

3.2.4 The genetic bottleneck may not be caused by the mucosal barrier *118*

3.2.5 Envelope proteins and transmissibility *118*

3.2.6 Comparisons with an animal model *119*

3.2.7 Aims *120*

3.3 MATERIALS AND METHODS 121

3.3.1 Participants *121*

3.3.2 Laboratory Methods *121*

3.3.3 Analysis methods *124*

3.3.4 Immunological profiling *128*

3.4 RESULTS 131

3.4.1 Case history of a single, third party exposure and recent seroconversion *131*

3.4.2 Sequences for phylogenetic analysis obtained from multiple viral genes *134*

3.4.3 Between-host phylogenetic analysis is consistent with the clinical history of a single donor *136*

3.4.4 Phylogenetic signal-to-noise analysis *138*

3.4.5 Phylogenetic analysis for sampled genes or fragments *139*

3.4.6 Comparison transmission recipient pair (R1 and R2) *160*

3.4.7 Differences in potential antigenic env variation between transmitted viruses in P1 and P2 ..... *162*

3.4.8 Immunological analysis *165*

3.5 DISCUSSION 170

3.5.1 Summary of key findings *170*

3.5.2 Sequences in P1 and P2 are highly-related *171*

3.5.3 There was no evidence that hard selection accounted for between-host divergence of variants *172*

3.5.4 Differences between different viral genes in terms of between-host as well as within-host genetic divergence *175*

3.6 CONCLUSION 179

3.6.1 Key Findings *179*

3.6.2 Concluding remarks *179*

CHAPTER 4 QUANTIFYING PRE-TRANSMISSION VIRAL DIVERGENCE 181

4.1 OVERVIEW 181

4.1.1 Key abbreviations *181*

4.1.2 Chapter summary and hypothesis *181*

4.2 INTRODUCTION 183

4.2.1 Quantifying transmissible diversity and implications for vaccine design *183*

4.2.2 Aims *183*

4.3 MATERIALS AND METHODS 184

4.3.1 Subjects *184*

4.3.2 Laboratory methods *184*

4.3.3 Coalescent analysis methods *184*

4.4 RESULTS 188

4.4.1 Summary of key points concerning between-host env tMRCA estimates *188*

4.4.2 Quantification of pre-transmission divergence in env *190*

4.4.3 Robust analysis in env C2V5 for estimating the tMRCA *192*

4.4.4 Robust analysis for tMRCA estimate substituting a larger env fragment *193*

4.4.5 Quantification of pre-transmission divergence in nef *193*

4.4.6 Quantification of pre-transmission divergence in pol *197*

4.4.7 Quantification of pre-transmission divergence in gag *200*

4.5 DISCUSSION 203

4.5.1 P1 and P2 were infected with very different env variants. *203*

4.5.2 Env fragments contained enough phylogenetic information for between-host tMRCA estimation using the GTR model *203*

4.5.3 Transmission of significant nef diversity *203*

4.5.4 P1 and P2 were infected with identical gag variants *204*

4.5.5 Evidence consistent with multi-variant pol transmission *205*

4.6 CONCLUSION 207

4.6.1 Key Findings *207*

4.6.2 Suggestions for further work *207*

CHAPTER 5 INFERRING INTRA-SUBTYPE HIV-1 MULTIPLE-INFECTION 209

5.1 OVERVIEW 209

5.1.1 Brief summary of background to chapter *209*

5.1.2 Brief summary of chapter and hypothesis *209*

5.2 INTRODUCTION 211

5.2.1 Multiple-infection *211*

5.2.2 Transmission, multiple-infection and the immune response *212*

5.2.3 Detection versus inference of multiple-infection within a subtype *212*

5.2.4 Points addressed in this chapter *214*

5.3 MATERIALS AND METHODS 215

5.3.1 Subjects *215*

5.3.2 Laboratory methods *215*

5.3.3 Phylogenetic Analysis *217*

5.3.4 Recombination Analysis *218*

5.4 RESULTS 220

5.4.1 Clinical evidence for exposure of P1 to a fourth host after day 63 ..... *220*

5.4.2 Sequence analysis *220*

5.4.3 Immunological analysis *227*

5.5 DISCUSSION 234

5.5.1 Identification of an epidemiologically unlinked, invading sequence *234*

5.5.2 Immunological implications of multiple-infection *235*

5.5.3 Comparison with other cases of intra-subtype multiple infection or superinfection *238*

5.6 CONCLUSION 240

5.6.1 Key Findings *240*

5.6.2 Further work to be done *240*

5.6.3 Concluding remarks *241*

CHAPTER 6 A QUALITATIVE DIPLOID MODEL OF MULTIPLE-INFECTION AND SUPERINFECTION 242

6.1 OVERVIEW 242

6.1.1 Brief summary of chapter and hypothesis *242*

6.2 INTRODUCTION 245

6.2.1 Overview *245*

6.2.2 Conflicting data predictions in previous approaches *246*

6.2.3 Qualitative approach to interpreting experimental data *248*

6.3 MODEL 250

6.3.1 Definitions and Assumptions *250*

6.3.2 The discrete deterministic model *251*

6.3.3 Discussion of simulation results in context of HIV-1 ..... *254*

6.3.4 Conclusions from the simple diploid model and simulations *258*

6.4 MATERIALS AND METHODS 259

6.4.1 Subjects *259*

6.4.2 Laboratory methods *259*

6.4.3 Phylogenetic Analysis *260*

6.4.4 Recombination Analysis *261*

6.4.5 Quantification of within-host diversification in terms of the tMRCA for the env sequences *262*

6.5 RESULTS 264

6.5.1 Two distinct clusters of sequences in env and gag sequence sampling *264*

6.5.2 Phylogenetic analysis *266*

6.5.3 Identification of recombinants in env and gag *272*

6.5.4 Quantification of within-host env diversity in terms of the tMRCA *274*

6.5.5 Immunological analysis *276*

6.6 DISCUSSION 281

6.6.1 Recombination demonstrated by SGA in env *281*

6.6.2 Superinfection by a dominant strain demonstrated in gag *284*

6.6.3 The qualitative diploid model allows predictions of HIV-1 evolution *286*

6.7 CONCLUSION 288

6.7.1 Future work *288*

6.7.2 Implications of my data for transmission and primary infection *288*

CHAPTER 7 DISCUSSION 289

7.1 OVERVIEW 289

7.2 KEY QUESTIONS REVISITED 290

7.2.1 Is there any evidence of hard selection of viral diversity at transmission? *290*

7.2.2 To what extent does recombination play a role in shaping viral diversity when two variants of the same subtype co-circulate at very different initial frequencies? *292*

7.3 FURTHER WORK AND FUTURE DIRECTIONS 298

7.3.1 Replication capacity comparison in multiple infection using the gag-protease region 298

7.3.2 Confirming the role of the mixed recombinant state in facilitating multiple-infection .298

7.4 CONCLUDING REMARKS 299

CHAPTER 8 APPENDIX 301

8.1 PCR PRIMERS 301

8.2 ADDITIONAL PRIMERS FOR SANGER SEQUENCING 303

8.3 POISSON FITTER HISTOGRAMS FOR CHAPTER 3 304

8.4 NEUTRALIZATION AND INFECTIVITY DATA FOR CHAPTER 3 307

8.5 ADDITIONAL ROBUST ANALYSIS FOR CHAPTER 4 309

8.5.1 Robust analysis in env C2V5 using the TN93 model 309

8.5.2 Robust analysis in env C2V5 using the HKY85 model 311

8.5.3 Rate heterogeneity among sites 313

8.6 ELISPOT DATA FOR SL9 MUTATION IN CHAPTERS 5 AND 6 315

CHAPTER 9 PUBLISHED PAPER 316

CHAPTER 10 REFERENCES 318

Figures

FIGURE 3.1 A. CD4+ COUNT FOR P1 VERSUS P2 OVER TIME 133

FIGURE 3.1 B. LOG VIRAL LOAD FOR P1 VERSUS P2 OVER TIME 133

FIGURE 3.1 C. KEY INFERRED EVENTS DURING THE FIRST 63 DAYS OF INFECTION 134

FIGURE 3.2 SAMPLED GENES OR FRAGMENTS (NOT DRAWN TO SCALE) 135

FIGURE 3.3 ZOOMED-IN PHYLOGENETIC TREES OF THE SMALL GAG FRAGMENT (A. AND B.) AND THE LARGE GAG FRAGMENT (C. AND D.) FOR P1 (BLUE) AND P2 (RED). ML TREES (A. AND C.) AND BAYESIAN MCMC BASED TREES (B. AND D.) ARE SHOWN. BOOTSTRAP VALUES (A. AND C.) OR POSTERIOR PROBABILITIES (B. AND D.) FOR THE CLUSTER ARE ALSO GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE (0.05 = 5%). 142

FIGURE 3.4 HIGHLIGHTER PLOT OF THE GAG FRAGMENT ALIGNMENTS FOR P1, SHOWING BASE SUBSTITUTIONS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: A=GREEN, T=RED, G=ORANGE, C=LIGHT BLUE, GAP=GREY, PINK CLOSED CIRCLE=APOBEC (A TO G), PINK OPEN DIAMOND=G TO A (ANTI-SENSE APOBEC). MOTIFS CONSISTENT WITH APOBEC SIGNATURES ARE MARKED IN PURPLE. 143

FIGURE 3.5 HIGHLIGHTER PLOT OF THE GAG FRAGMENT ALIGNMENTS FOR P1, SHOWING NON-SYNONYMOUS VERSUS SYNONYMOUS MUTATIONS. BASE POSITION ALONG THE SEQUENCE

ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: SYNONYMOUS=GREEN, NON-SYNONYMOUS=RED. 144

FIGURE 3.6 HIGHLIGHTER PLOT OF THE GAG FRAGMENT ALIGNMENTS FOR P2, SHOWING BASE SUBSTITUTIONS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: A=GREEN, T=RED, G=ORANGE, C=LIGHT BLUE, GAP=GREY, PINK CLOSED CIRCLE=APOBEC (A TO G), PINK OPEN DIAMOND=G TO A (ANTI-SENSE APOBEC). MOTIFS CONSISTENT WITH APOBEC SIGNATURES ARE MARKED IN PURPLE. 145

FIGURE 3.7 HIGHLIGHTER PLOT OF THE GAG FRAGMENT ALIGNMENTS FOR P1, SHOWING NON-SYNONYMOUS VERSUS SYNONYMOUS MUTATIONS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: SYNONYMOUS=GREEN, NON-SYNONYMOUS=RED. 146

FIGURE 3.8 ZOOMED-IN ML (A.) AND BAYESIAN MCMC BASED (B.) TREES FOR GAG. THE REGIONS OF PROBABLE INTER-MINGLING BETWEEN P1 (BLUE) AND P2 (RED) SEQUENCES ARE HIGHLIGHTED IN GREEN DASHED BOXES. THE SCALE BAR SHOWS GENETIC DISTANCE, SO THAT 0.05 = 5%. 147

FIGURE 3.9 ZOOMED-IN PHYLOGENETIC TREES OF THE LARGE ENV FRAGMENT (A. AND B.) AND THE ENV C2V5 FRAGMENT (C. AND D.) FOR P1 (BLUE) AND P2 (RED). ML TREES (A. AND C.) AND BAYESIAN MCMC BASED TREES (B. AND D.) ARE SHOWN. BOOTSTRAP VALUES (A. AND C.) OR POSTERIOR PROBABILITIES (B. AND D.) FOR THE CLUSTER ARE ALSO GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE (0.05 = 5%).....148

FIGURE 3.10 HIGHLIGHTER PLOT OF BASE MISMATCHES IN THE UNSTRIPPED GP120 REGION RELATIVE TO THE P1 GP120 DAY 63 SGA CONSENSUS. THE X-AXIS INDICATES BASE POSITION ALONG THE SEQUENCE ALIGNMENT. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS ARE: A=GREEN, T=RED, G=ORANGE, C=LIGHT BLUE, GAP=GREY. 149

FIGURE 3.11 HIGHLIGHTER PLOT OF UNSTRIPPED GP120 REGION SHOWING BASE MISMATCHES RELATIVE TO THE DAY 63 P2 SGA CONSENSUS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: A=GREEN, T=RED, G=ORANGE, C=LIGHT BLUE, GAP=GREY, PINK CLOSED CIRCLE=APOBEC (A TO G), PINK OPEN DIAMOND=G TO A (ANTI-SENSE APOBEC). SEQUENCES SHARING A COMMON G-TO-A SUBSTITUTION ARE INDICATED BY SEQUENCE NAMES IN PURPLE FONT. SEQUENCES SHARING A COMMON T-TO-A SUBSTITUTION ARE INDICATED BY SEQUENCE NAMES IN GREY BOXES.....150

FIGURE 3.12 HIGHLIGHTER PLOT OF UNSTRIPPED GP120 REGION SHOWING SYNONYMOUS (GREEN) AND NON-SYNONYMOUS (RED) SUBSTITUTIONS RELATIVE TO THE P2 DAY 63 SGA CONSENSUS.

THE X-AXIS INDICATES CODON POSITION ALONG THE SEQUENCE ALIGNMENT. THE Y-AXIS INDICATES SEQUENCE NAMES. THE COLOUR KEY IS GIVEN TO THE LEFT OF THE LEFT Y-AXIS. SEQUENCES SHARING A COMMON G-TO-A NON-SYNONYMOUS SUBSTITUTION ARE INDICATED BY SEQUENCE NAMES IN PURPLE FONT. SEQUENCES SHARING A COMMON T-TO-A NON-SYNONYMOUS SUBSTITUTION ARE INDICATED BY SEQUENCE NAMES IN GREY BOXES.....151

FIGURE 3.13 ZOOMED-IN PHYLOGENETIC TREES OF THE NEF-ONLY FRAGMENT (A. AND B.) AND THE LARGE NEF-CONTAINING FRAGMENT (C. AND D.) FOR P1 (BLUE) AND P2 (RED). ML TREES (A. AND C.) AND BAYESIAN MCMC BASED TREES (B. AND D.) ARE SHOWN. BOOTSTRAP VALUES (A. AND C.) OR POSTERIOR PROBABILITIES (B. AND D.) FOR THE CLUSTER ARE GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE (0.05 = 5%). 152

FIGURE 3.14 HIGHLIGHTER PLOT OF THE NEF-ONLY FRAGMENT SHOWING.....153

SYNONYMOUS VERSUS NON-SYNONYMOUS MUTATIONS RELATIVE TO THE P2 DAY 63 CLONAL SEQUENCE CONSENSUS. CODON POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: SYNONYMOUS=GREEN, NON-SYNONYMOUS=RED. SEQUENCES WITH A NON-SYNONYMOUS MUTATION IN AT LEAST ONE CODON POSITION IN COMMON ARE INDICATED BY SEQUENCE NAMES IN PURPLE FONT. 153

FIGURE 3.15 HIGHLIGHTER PLOT OF THE NEF-ONLY FRAGMENT SHOWING BASE MISMATCHES RELATIVE TO THE P2 DAY 63 CLONAL SEQUENCE CONSENSUS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED BY THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. COLOURS: SYNONYMOUS=GREEN, NON-SYNONYMOUS=RED. SEQUENCES WITH A NONSYNONYMOUS MUTATION IN AT ONE CODON POSITION IN COMMON ARE INDICATED BY SEQUENCE NAMES IN PURPLE FONT. 154

FIGURE 3.16 HIGHLIGHTER PLOT FOR P1 POL FRAGMENT BASE MISMATCHES RELATIVE TO THE P1 DAY 63 CLONAL SEQUENCE CONSENSUS. BASE POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED ON THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. THE NAMES OF CLONAL SEQUENCES REPRESENTING SUSPECTED MINOR VARIANTS ARE INDICATED BY PURPLE TEXT. COLOURS: A=GREEN, T=RED, G=ORANGE, C=LIGHT BLUE, GAP=GREY. 157

FIGURE 3.17 HIGHLIGHTER PLOT FOR SYNONYMOUS VERSUS NONSYNONYMOUS SUBSTITUTIONS IN THE P2 POL FRAGMENT RELATIVE TO THE P2 DAY 63 CLONAL SEQUENCE CONSENSUS. CODON POSITION ALONG THE SEQUENCE ALIGNMENT IS INDICATED ON THE X-AXIS. SEQUENCE NAMES ARE INDICATED ON THE Y-AXIS. THE NAMES OF CLONAL SEQUENCES REPRESENTING SUSPECTED MINOR VARIANTS ARE INDICATED BY PURPLE TEXT. COLOURS: SYNONYMOUS=GREEN, NON-SYNONYMOUS=RED. 157

FIGURE 3.18 ZOOMED-IN ML (A.) AND BAYESIAN MCMC (B.) TREE FOR POL CLONAL SEQUENCES

SHOWING (A.) BOOTSTRAP OR (B.) POSTERIOR PROBABILITY VALUES FOR INTERNAL BRANCHES (TERMINATING AT NODES INDICATED BY BLACK DOTS) FOR P1 (BLUE DOTS) AND P2 (RED DOTS). SUSPECTED MINORITY VARIANTS HIGHLIGHTED IN COLOURED DASHED BOXES FOR P1 (PURPLE) AND P2 (GREEN). THE SCALE BAR SHOWS GENETIC DISTANCE, SO THAT 0.05 = 5%. 160

FIGURE 3.19 ML TREES FOR A. GAG AND B. ENV IN R1 (BLUE) AND R2 (RED). BOOTSTRAP SUPPORT FOR THE CLUSTER IS ALSO GIVEN AS A PERCENTAGE. THE SCALE BAR SHOWS GENETIC DISTANCE, SO THAT 0.05 = 5%. 162

FIGURE 3.20 PNLGSS ALONG SGA ENV GP120 AA SEQUENCES FROM P1 (BLUE) AND P2 (RED). CODON POSITION IS INDICATED ALONG THE X-AXIS. THE FRACTION OF SGA SEQUENCES WITH A PNLGS AT A POSITION IS INDICATED BY THE Y-AXIS. REGIONS OF INTEREST ARE INDICATED BY YELLOW BOXES. POSITIONS WHERE 100% OF SEQUENCES IN BOTH P1 AND P2 HAVE A PNLGS ARE INDICATED BY SMALL STARS ABOVE THE GRAPH. 163

FIGURE 4.1 ESTIMATION OF PRE-TRANSMISSION DIVERGENCE IN ENV C2V5: GTR MODEL, ULRM CLOCK, NORMAL MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 191

FIGURE 4.2 ENV C2V5: GTR MODEL, STRICT MOLECULAR CLOCK; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 192

FIGURE 4.3 WHOLE ENV FRAGMENT (GP160 CODING START TO GP120 END): GTR MODEL, STRICT MOLECULAR CLOCK, NO MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 193

FIGURE 4.4 A TMRCA ANALYSIS FOR NEF: TN93 MODEL, ULRM CLOCK; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 196

FIGURE 5.1 REGIONS OF THE HIV-1 GENOME ANALYSED IN CHAPTERS 5 AND 6 216

FIGURE 5.2 THE A. FULL ML TREE FOR NEF POPULATION AMPLIFICATION SEQUENCES AND ZOOMED IN VIEWS OF THE B. ORIGINAL SEQUENCES CLUSTER (N1) AND C. THE NEW P1 SEQUENCES CLUSTER (N2). FOR P1, THE COLOURS ARE: INTERMEDIATE BLUE=DAY 63, GREY=DAY 137, DARK BLUE=DAY 235, BROWN=DAY 474, GREEN=DAY 819. FOR P2, THE COLOURS ARE: LIGHT RED=DAY 63, PINK=DAY 137, DARK RED=DAY 235. BOOTSTRAP SUPPORT FOR CLUSTERS IN SHOWN AS A PERCENTAGE. THE SCALE BAR REPRESENTS GENETIC DISTANCE, SO THAT 0.05=5%. THE NEAREST BRANCHES WITH THE NEAREST SEQUENCES THAT ARE NEITHER FROM P1 NOR P2 ARE SHOWN FOR COMPARISON. 221

FIGURE 5.3 ML TREE FOR ENV POPULATION AMPLIFICATION SEQUENCES. FOR P1, THE COLOURS ARE: INTERMEDIATE BLUE=DAY 63, DARK BLUE=DAY 235, PURPLE=DAY 439, BROWN=474 AND GREEN=DAY 819. FOR P2, THE COLOURS ARE: LIGHT RED=DAY 63 AND DARK RED=DAY 235. BOOTSTRAP SUPPORT FOR CLUSTERS IN SHOWN AS A PERCENTAGE. THE SCALE BAR

REPRESENTS GENETIC DISTANCE, SO THAT 0.05=5%. 222

FIGURE 5.4 ML TREE FOR GAG POPULATION AMPLIFICATION SEQUENCES. BOOTSTRAP SUPPORT FOR CLUSTERS ARE SHOWN AS PERCENTAGES. THE SCALE BAR REPRESENTS GENETIC DISTANCE, SO THAT 0.05=5%. CLUSTER 1 AND CLUSTER 2 ARE INDICATED. ARROWS POINT TOP THE SUSPECTED RECOMBINANT SEQUENCES. FOR P1, THE COLOURS ARE: INTERMEDIATE BLUE=DAY 63, GREY=DAY 137, LIGHT BLUE=DAY 185, DARK BLUE=DAY 235, DARK GREEN= DAY 397, PURPLE=DAY 439, BROWN=DAY 474, GREEN=DAY 819. FOR P2, THE COLOURS ARE: LIGHT RED=DAY 63, PINK=DAY 137, DARK RED=DAY 235. 223

FIGURE 5.5 HIGHLIGHTER PLOT OF GAG. THE NAMES OF SUSPECTED RECOMBINANT SEQUENCES ARE HIGHLIGHTED IN BLUE FONT. THE NAME MOST DIVERGENT SEQUENCE, USED AS THE SURROGATE REFERENCE FOR THE INVADING SEQUENCE, IS HIGHLIGHTED IN PURPLE. THE NAMES OF EXAMPLE CLUSTER 1 SEQUENCES FROM DAYS 63 AND 235 ARE HIGHLIGHTED IN GREEN AND RED FONT, RESPECTIVELY. 225

FIGURE 5.6 QUERY-VERSUS-REFERENCE PLOTS FOR THE SUSPECTED RECOMBINANT SEQUENCE P1 DAY 235 GAG 04 OF A. THE FRACTION OF SEQUENCE SIMILARITY (OUT OF 1.0) AND B. THE PERCENTAGE OF PERMUTED TREES THAT SUPPORT CLUSTERING WITH THE REFERENCE SEQUENCE. THE GREEN LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 63 GAG 01. THE RED LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 235 GAG 01. THE PURPLE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 439 GAG 01. 226

FIGURE 5.7 QUERY-VERSUS-REFERENCE PLOTS FOR THE SUSPECTED RECOMBINANT SEQUENCE P1 DAY 439 GAG 03 OF A. THE FRACTION OF SEQUENCE SIMILARITY (OUT OF 1.0) AND B. THE PERCENTAGE OF PERMUTED TREES THAT SUPPORT CLUSTERING WITH THE REFERENCE SEQUENCE. THE GREEN LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 63 GAG 01. THE RED LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 235 GAG 01. THE PURPLE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 439 GAG 01. 227

FIGURE 6.1 NO MULTIPLE-INFECTION OR SUPER-INFECTION. RELATIVE FREQUENCIES OF GENETIC SUBUNITS FROM PHYLOGENY A (BLUE) AND PHYLOGENY B (RED) ARE SHOWN. THE HORIZONTAL AXIS IS TIME IN GENERATIONS (T) AND THE VERTICAL ACCESS IS RELATIVE FREQUENCY (0 TO 1.0). 255

FIGURE 6.2 SUPERINFECTION BY PHYLOGENY B. RELATIVE FREQUENCIES OF GENETIC SUBUNITS FROM PHYLOGENY A (BLUE) AND PHYLOGENY B (RED) ARE SHOWN. THE HORIZONTAL AXIS IS TIME IN GENERATIONS (T) AND THE VERTICAL ACCESS IS RELATIVE FREQUENCY (0 TO 1.0). ..256

FIGURE 6.3 MULTIPLE-INFECTION WHEN THE POPULATION MOSTLY RESEMBLES THE ORIGINAL STRAIN. FREQUENCIES OF GENETIC SUBUNITS FROM PHYLOGENY A (BLUE) AND PHYLOGENY B (RED) ARE SHOWN. THE HORIZONTAL AXIS IS TIME IN GENERATIONS (T) AND THE VERTICAL

AXIS IS RELATIVE FREQUENCY (0 TO 1.0).....257

FIGURE 6.4 MULTIPLE-INFECTION WHEN THE POPULATION MOSTLY RESEMBLES THE INVADING STRAIN. FREQUENCIES OF GENETIC SUBUNITS FROM PHYLOGENY A (BLUE) AND PHYLOGENY B (RED) WHEN THE RESULT IS STABLE MULTIPLE-INFECTION WHERE THE POPULATION MORE FREQUENTLY RESEMBLE PHYLOGENY B. THE HORIZONTAL AXIS IS TIME IN GENERATIONS (T) AND THE Y ACCESS IS RELATIVE FREQUENCY (0 TO 1.0). 258

FIGURE 6.5 ZOOMED-IN VIEWS OF A. ML AND B. BAYESIAN MCMC BASED TREES OF THE SGA LONG ENV FRAGMENT SEQUENCES AT ALL SAMPLED TIME POINTS FOR P1 AND P2. P1 COLOURS: INTERMEDIATE BLUE=DAY 63, DARK BLUE=DAY 235, LIGHT BROWN=DAY 474. P2 COLOURS: LIGHT RED=DAY 63, DARK RED=DAY 235. BOOTSTRAP VALUES FOR CLUSTER SUPPORT AT INTERNAL NODES ARE GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE, SO THAT 0.05=5%. FEWER REFERENCE SEQUENCES WERE USED TO CONSTRUCT THE BAYESIAN MCMC BASED TREE (71 FOR BAYESIAN MCMC VERSUS 147 FOR ML).....268

FIGURE 6.6 (PREVIOUS PAGE) ZOOMED-IN ML TREE OF CLONAL GAG SEQUENCES FROM P1 AND P2. COLOURED LABELS IDENTIFY SEQUENCES FROM DIFFERENT PATIENTS AND DIFFERENT TIME POINTS. FOR P1, THE COLOURS ARE: INTERMEDIATE BLUE=DAY 63, GREY=DAY 137, LIGHT BLUE=DAY 185, DARK BLUE=DAY 235, PURPLE=DAY 439, BROWN=474, GREEN=DAY 819. FOR P2, THE COLOURS ARE: LIGHT RED=DAY 63, PINK=DAY 137, DARK RED=DAY 235. BOOTSTRAP VALUES FOR CLUSTER SUPPORT ARE GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE (0.05 = 5%). 270

FIGURE 6.7 (PREVIOUS PAGE) ZOOMED-IN BAYESIAN MCMC BASED TREE OF CLONAL GAG SEQUENCES FROM P1 AND P2. COLOURED LABELS IDENTIFY SEQUENCES FROM DIFFERENT PATIENTS AND DIFFERENT TIME POINTS. FOR P1, THE COLOURS ARE: INTERMEDIATE BLUE=DAY 63, GREY=DAY 137, LIGHT BLUE=DAY 185, DARK BLUE=DAY 235, PURPLE=DAY 439, BROWN=474, GREEN=DAY 819. FOR P2, THE COLOURS ARE: LIGHT RED=DAY 63, PINK=DAY 137, DARK RED=DAY 235. BOOTSTRAP VALUES FOR CLUSTER SUPPORT ARE GIVEN AS PERCENTAGES. THE SCALE BAR SHOWS GENETIC DISTANCE (0.05 = 5%). CLUSTER 2 IS INDICATED. 272

FIGURE 6.8 QUERY-VERSUS-REFERENCE PLOTS FOR THE SUSPECTED RECOMBINANT SEQUENCE P1 DAY 235 ENV SGA 13. PLOTS SHOW (TOP) THE FRACTION OF SEQUENCE SIMILARITY (OUT OF 1.0) AND (BOTTOM) THE PERCENTAGE OF PERMUTED TREES THAT SUPPORT CLUSTERING WITH THE REFERENCE SEQUENCE. THE RED LINE REPRESENTS THE REFERENCE CONSENSUS SEQUENCE FOR P1 AT DAY 63. THE BLUE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 235 ENV SGA 14. THE PURPLE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 439 GAG 01. THE GREY LINE REPRESENTS THE REFERENCE CONSENSUS SEQUENCE FOR P2 AT DAY 63.....273

FIGURE 6.9 QUERY-VERSUS-REFERENCE PLOTS FOR THE SUSPECTED RECOMBINANT SEQUENCE P1 DAY 474 ENV SGA 1. PLOTS SHOW (TOP) THE FRACTION OF SEQUENCE SIMILARITY (OUT OF 1.0) AND (BOTTOM) THE PERCENTAGE OF PERMUTED TREES THAT SUPPORT CLUSTERING WITH THE REFERENCE SEQUENCE. THE RED LINE REPRESENTS THE REFERENCE CONSENSUS SEQUENCE FOR P1 AT DAY 63. THE BLUE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 235 ENV SGA 14. THE PURPLE LINE REPRESENTS THE REFERENCE SEQUENCE P1 DAY 439 GAG 01. THE GREY LINE REPRESENTS THE REFERENCE CONSENSUS SEQUENCE FOR P2 AT DAY 63. THE PURPLE LINE REPRESENTS THE REFERENCE CONSENSUS SEQUENCE FOR P1 AT DAY 235. 274

FIGURE 8.1 HISTOGRAMS FOR THE HD FREQUENCY DISTRIBUTION FOR THE LONG ENV FRAGMENT IN P1 (A. AND B.) AND P2 (C. AND D.), WITH APOBEC SIGNATURES REMOVED. THE X-AXIS PLOTS HAMMING DISTANCE VALUES FROM THE WITHIN-HOST CONSENSUS FOR EACH RECIPIENT. THE Y-AXIS PLOTS HD FREQUENCY COUNTS AS A PERCENTAGE OF SAMPLES SEQUENCED. THE RED LINE IN A. AND C. PLOTS THE EXPECTED HD FREQUENCY COUNTS UNDER A POISSON DISTRIBUTION MODEL. THE RED LINES IN B. AND D. PLOT THE EXPECTED HD FREQUENCY COUNTS IF THE DATA FOLLOWED A STAR-LIKE PHYLOGENY UNDER THE SELF-CONVOLUTION MODEL IMPLEMENTED IN POISSON-FITTER. 304

FIGURE 8.2 HISTOGRAMS FOR THE HD FREQUENCY DISTRIBUTION FOR THE LONG ENV FRAGMENT IN P2, WITH APOBEC SIGNATURES INCLUDED. THE X-AXIS PLOTS HAMMING DISTANCE VALUES FROM THE WITHIN-HOST CONSENSUS FOR EACH RECIPIENT. THE Y-AXIS PLOTS HD FREQUENCY COUNTS AS A PERCENTAGE OF SAMPLES SEQUENCED. THE RED LINE IN THE TOP PLOT IS THE EXPECTED HD FREQUENCY COUNTS UNDER A POISSON DISTRIBUTION MODEL. THE RED LINES IN THE BOTTOM PLOT IS THE EXPECTED HD FREQUENCY COUNTS IF THE DATA FOLLOWED A STAR-LIKE PHYLOGENY UNDER THE SELF-CONVOLUTION MODEL IMPLEMENTED IN POISSON-FITTER. 305

FIGURE 8.3 HISTOGRAMS FOR THE HD FREQUENCY DISTRIBUTION FOR THE ENV C2V5 FRAGMENT IN P2, WITH APOBEC SIGNATURES INCLUDED. THE X-AXIS PLOTS HAMMING DISTANCE VALUES FROM THE WITHIN-HOST CONSENSUS FOR EACH RECIPIENT. THE Y-AXIS PLOTS HD FREQUENCY COUNTS AS A PERCENTAGE OF SAMPLES SEQUENCED. THE RED LINE IN THE TOP PLOT IS THE EXPECTED HD FREQUENCY COUNTS UNDER A POISSON DISTRIBUTION MODEL. THE RED LINES IN THE BOTTOM PLOT IS THE EXPECTED HD FREQUENCY COUNTS IF THE DATA FOLLOWED A STAR-LIKE PHYLOGENY UNDER THE SELF-CONVOLUTION MODEL IMPLEMENTED IN POISSON-FITTER. 306

FIGURE 8.4 IN VITRO NEUTRALIZATION OF PARTICIPANT ENV-PSEUDOTYPED VIRUSES BY AUTOLOGOUS (SAME PARTICIPANT) AND HETEROLOGOUS (PARTICIPANT'S PARTNER) SERUM FOR EACH OF THE CLONES TESTED. DATA FROM THE MOST CONCENTRATED SERUM DILUTIONS

ARE SHOWN, WITH THE COLOUR SCHEME PROVIDED BY A LEGEND AT THE RIGHT OF EACH GRAPH. 307

FIGURE 8.5 INFECTIVITY OF THE ENV-PSEUDOTYPED VIRUS FROM EACH PARTICIPANT. DATA FOR DIFFERENT FOLD DILUTIONS OF PSEUDOVIRUS AFTER NORMALIZATION TO REVERSE TRANSCRIPTASE LEVELS. A COLOUR KEY INDICATES THE DILUTIONS TESTED.....308

FIGURE 8.6 ENV C2V5: TN93 MODEL, ULRM CLOCK, NO MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 309

FIGURE 8.7 ENV C2V5: TN93 MODEL, UNCORRELATED LOGNORMAL MOLECULAR CLOCK, NORMAL MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2.....310

FIGURE 8.8 ENV C2V5: TN93 MODEL, STRICT MOLECULAR CLOCK, NO SUBSTITUTION RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2.....311

8.9 ENV C2V5: HKY85 MODEL, ULRM CLOCK, NO MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 312

FIGURE 8.10 ENV C2V5: HKY85 MODEL, ULRM CLOCK, NORMAL MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 313

FIGURE 8.11 TN93+Γ MODEL, ULRM CLOCK, NO MEAN RATE PRIOR; BLUE CLOSED DOTS=P1, RED CLOSED DOTS=P2. 314

Tables

TABLE 1.1 SUMMARY OF KEY DEFINITIONS 71

TABLE 1.2 SUMMARY OF EXPERIMENTAL DATA 84

TABLE 2.1 ABBREVIATIONS....87

TABLE 2.2 SUMMARY OF COHORTS 87

TABLE 2.3 SUMMARY OF FIEBIG STAGE DEFINITIONS FROM THE ONSET OF DETECTABLE VIRAEMIA (DAY 0). (FIEBIG ET AL. 2003; FIEBIG ET AL. 2005) 89

TABLE 2.4 ESTIMATED TIMING AND DURATION OF FIEBIG STAGES FROM THE ONSET OF DETECTABLE VIRAEMIA (DAY 0). (FIEBIG ET AL. 2003; FIEBIG ET AL. 2005) 89

TABLE 2.5 ABBREVIATIONS....93

TABLE 2.6 ABBREVIATIONS....96

TABLE 3.1 ABBREVIATIONS....116

TABLE 3.2 SUMMARY OF ML WITHIN- AND BETWEEN-HOST DISTANCE 140

TABLE 3.3 CODON SUBSTITUTION PATTERNS FOR HXB2, THE P1 AND P2 CONSENSUS SEQUENCES, AND EIGHT P2 CLONES AT THREE CODON POSITIONS. DIFFERENCES FROM THE P2 CONSENSUS (**) ARE SHOWN IN BOLD AND UNDERLINED. CODON POSITION IS NUMBERED AS THE HXB2

POSITION. NUCLEOTIDES AND AMINO ACIDS ARE NAMED ACCORDING TO THE SINGLE-LETTER CODE. 156

TABLE 3.4 CODON SUBSTITUTION PATTERNS FOR HXB2, THE CONSENSUS SEQUENCE FOR EACH RECIPIENT AND THE FIVE SUSPECTED MINOR VARIANT SEQUENCES IN P2. DIFFERENCES FROM THE P2 CONSENSUS (**) ARE SHOWN IN BOLD AND UNDERLINED. 159

TABLE 4.1 ABBREVIATIONS....181

TABLE 4.2 PRE-TRANSMISSION TMRCA ESTIMATES FOR ALL MODELS USED TO ANALYSE ENV C2V5 189

TABLE 4.3 PRE-TRANSMISSION TMRCA ESTIMATES FOR ALL MODELS USED TO ANALYSE NEF....195

TABLE 4.4 MEAN RATE ESTIMATES FOR EQUIVALENT MODELS IN NEF AND ENV C2V5 196

TABLE 4.5 PRE-TRANSMISSION TMRCA ESTIMATES FOR ALL MODELS USED TO ANALYSE THE POL FRAGMENT 197

TABLE 4.6 MEAN RATE ESTIMATES FOR EQUIVALENT MODELS IN ENV C2V5, NEF AND THE POL FRAGMENT. 200

TABLE 4.7 PRE-TRANSMISSION TMRCA ESTIMATES FOR ALL MODELS USED TO ANALYSE THE GAG FRAGMENT 201

TABLE 4.8 MEAN RATE ESTIMATES FOR EQUIVALENT MODELS IN THE GAG FRAGMENT AND ENV C2V5....201

TABLE 5.1 TIME POINTS SAMPLED FROM P1 215

TABLE 5.2 ELISPOT RESPONSES FOR P1 IN THE REGIONS SEQUENCED. THE AMINO ACID SEQUENCES FOR EACH SPARTAC ELISPOT PEPTIDE CODE ARE GIVEN. THE HXB2 AMINO ACID LOCATION (AA) IS GIVEN AND MUTATIONS AWAY FROM THE SPARTAC ELISPOT PEPTIDE SEQUENCE ARE HIGHLIGHTED IN RED, BOLD AND UNDERLINED FONT. THE ELISPOT RESPONSE UNITS ARE SPOT FIRMING UNITS (SFU) PER MILLION CELLS. OLP = OVERLAPPING PEPTIDE, OP = OPTIMAL PEPTIDES. DATE: P1 RESPONSE GIVES THE DAY (E.G. D63 = DAY 63) WHEN A POSITIVE RESPONSE WAS DETECTED AND ITS MAGNITUDE (MINUS THE BACKGROUND AND 50 SFU PER MILLION CELLS). THE MATCHED CLASS I HLA IS GIVEN FOR THE OPTIMAL AND VARIANT PEPTIDES. MINOR VARIANTS SEEN ONLY IN INFERRED RECOMBINANTS IN P1 CLUSTER 2 ARE SHOWN IN BLACK BOLD. 231

TABLE 5.3 OTHER PUBLISHED CASES OF INTRA-SUBTYPE MULTIPLE-INFECTION OR SUPERINFECTION. THE NUMBER OF PARTICIPANTS (N) INFECTED WITH MORE THAN ONE INTRA-SUBTYPE STRAIN, THE SUBTYPE AND WHETHER MULTIPLE-INFECTION OR SUPERINFECTION WAS DETECTED. ONLY PARTICIPANTS INFECTED WITH THE SAME SUBTYPE HAVE BEEN INCLUDED. 239

TABLE 6.1 THE NUMBER OF PRIMER SETS REPRESENTED IN THE CLONAL GAG SAMPLING FROM EACH TIME POINT FOR P1 AND P2 . CLONING WAS UNSUCCESSFUL AFTER TWO ATTEMPTS FOR THE

TIME POINTS IN RED. 259

TABLE 6.2 THE TIME POINTS WHERE ENV SINGLE GENOME AMPLIFICATION (SGA) WAS SUCCESSFUL WITH A SINGLE PRIMER SET ARE INDICATED IN BLACK. THE TIME POINT WHERE SGA WAS SUCCESSFUL WITH A SINGLE PRIMER SET BUT WAS ATTEMPTED WITH TWO PRIMER SETS DUE TO LOW SEQUENCE SAMPLE NUMBER IS INDICATED IN GREEN. THE TIME POINT WHERE SGA WAS UNSUCCESSFUL FOLLOWING ATTEMPTS WITH TWO PRIMER SETS ARE INDICATED IN RED. TIME POINTS WHERE SGA WAS NOT ATTEMPTED ARE INDICATED IN BLUE. 260

TABLE 6.3 THE NUMBER OF PARTIAL- OR FULL-FRAGMENT SEQUENCES SAMPLED FOR EACH GENE/TECHNIQUE AT EACH TIME POINT. BRACKETED NUMBERS ARE THE NUMBER OF CLONES OBTAINED WITH EACH PRIMER SET. 265

TABLE 6.4 NUMBER OF PNLGSS AT DIFFERENT TIME POINTS IN THE V1V4 REGION OF ENV GP120.278

TABLE 6.5 ALIGNED AA SEQUENCES, FROM THE BEGINNING OF THE C3 REGION UNTIL HXB2 ENV AA POSITION 376. SEQUENCE NAME ABBREVIATIONS ARE: $P1_{CON.}$=CONSENSUS SEQUENCE FOR P1, $P2_{CON.}$=CONSENSUS SEQUENCE FOR P2, $P1_{SGA}$=SGA SEQUENCE FROM P1. MUTATIONS ARE HIGHLIGHTED IN RED AND UNDERLINED, RELATIVE TO THE P1 CONSENSUS SEQUENCE FROM DAY 63, WHICH IS HIGHLIGHTED IN BLACK BOLD. THE NAMES OF THE TWO MOST DIVERGENT SGA SEQUENCES IN P1 AT DAY 235 ARE HIGHLIGHTED IN BLUE. THE HXB2 SUBTYPE B REFERENCE SEQUENCE IS GIVEN FOR COMPARISON. 278

TABLE 6.6 ALIGNED SEQUENCES FOR THE ENV V4 REGION (HXB2 ENV AA 385 TO 418). SEQUENCE NAME ABBREVIATIONS ARE: $P1_{CON.}$=CONSENSUS SEQUENCE FOR P1, $P2_{CON.}$=CONSENSUS SEQUENCE FOR P2, $P1_{SGA}$=SGA SEQUENCE FROM P1. MUTATIONS ARE HIGHLIGHTED IN RED AND UNDERLINED, RELATIVE TO THE P1 CONSENSUS SEQUENCE FROM DAY 63, WHICH IS HIGHLIGHTED IN BLACK BOLD. THE NAMES OF THE TWO MOST DIVERGENT SGA SEQUENCES IN P1 AT DAY 235 ARE HIGHLIGHTED IN BLUE. THE HXB2 SUBTYPE B REFERENCE SEQUENCE IS GIVEN FOR COMPARISON. 279

TABLE 8.1 PCR PRIMERS. THE GENE FROM WHICH IN WHICH THE TARGET FRAGMENT SITS IS GIVEN IN THE FIRST COLUMN. 301

TABLE 8.2 ADDITIONAL PRIMERS FOR SANGER SEQUENCING 303

TABLE 8.3 EPITOPE SEQUENCES AND INTERFERON- γ ELISPOT RESPONSES TO SL9 WILD-TYPE PEPTIDE, SL9 VARIANT PEPTIDES ASSOCIATED WITH ESCAPE FROM HLA-A*0201, AND OVERLAPPING PEPTIDES (OLP) CONTAINING PART OR ALL OF SL9. A POSITIVE RESPONSE (+) WAS DEFINED AS GREATER THAN 50 SPOTS/MILLION CELLS (MINUS BACKGROUND SPOTS OBSERVED IN THE NEGATIVE CONTROL). THE WILD-TYPE (WT) PEPTIDE IS HIGHLIGHTED IN BOLD BLACK FONT. THE SL9 Y79F MUTANT FOUND IN P1 FROM DAY 235 IS HIGHLIGHTED IN BLUE FONT. AMINO ACID SUBSTITUTIONS AWAY FROM THE WT ARE HIGHLIGHTED IN RED,

BOLD AND UNDERLINED FONT.....315

Abstract

The central problem for researchers of HIV-1 evolution is explaining the apparent design of the virus for causing pandemic infection in humans: understanding how HIV-1 spreads is key to halting the pandemic. Current knowledge of how HIV-1 spreads from host to host is based upon experimental observation and indirect inferences informed by theory. The hypothesis of this thesis is that diversity of HIV-1 around the time of transmission is important for viral adaptation to a new human host, rather than intrinsic superiority of particular strains found in infectious fluids from human donor hosts, and that studying recombination is important for understanding this behaviour. To demonstrate the apparent randomness of transmission, I test the null-hypothesis that hard selection accounts for between-host viral divergence in a rare case study of contemporaneous infection. I explain how the experimental data that I have generated and the analyses I have carried out address certain basic assumptions and predictions about HIV-1 transmission and may inform current strategies for vaccine design. Specifically, my approach contributes to the current literature on HIV-1, by investigating an alternative hypothesis to the single virion theory of sexual transmission and by characterizing the role of recombination in a pseudodiploid virus following multiple-infection.

Statement of Authorship

Clinical data were provided by Dr Sarah Fidler and Mr Scott Mullaney at Imperial College, London, and Dr Fabian Chen at the Royal Berkshire Hospital, Reading.

Population amplification and sequencing of reference sequences, from SPARTAC and the St Mary's Open Study, in non-*env* genes was carried out by Dr Anna Duda, Ms Laura Lee-Turner, Ms Nicola Robinson, Ms Helen Brown and Mr Peter Flanagan at the Sir Peter Medawar Building for Pathogen Research, University of Oxford.

Clonal sequences for *pol* and the short fragment of *gag* presented in Chapter 3 were generated by Dr Anna Duda.

Clonal *gag* sequences for the two participants from the Royal Berkshire Hospital in Chapter 3 were generated by Mr Peter Flanagan.

All ELISpots for SPARTAC participants were carried out by Ms Nicola Robinson and Ms Helen Brown.

Env-pseudotyped viruses and neutralization assay data were generated by Dr David Bonsall in Professor Myra McClure's laboratory at Imperial College, London. Dr Bonsall also generated the figures shown in the Appendix (Chapter 8).

Acknowledgments

I gratefully acknowledge the support of my supervisors, Professor Rodney Phillips and Professor Paul Klenerman, without whom this thesis would not have been possible. I am also indebted to Dr Aris Katzourakis, Dr Claire El Mouden, Dr Rebecca Payne, Dr Julia Garcia-Prado, Dr Gary Huang, Dr Hayley Crawford and Dr John Frater, for their patience and interesting discussions. I have also benefited from the laboratory supervision or technical assistance of Helen Brown, Nicola Robinson and Dr Nicole Yager. I would also like to thank my collaborators at Imperial College, London, especially Dr Sarah Fidler and Dr David Bonsall. I would like to thank my partner, Mr Edward Hancock, for his patience and support. I would also like to thank my brother, Mr David English, for his invaluable computer skills and hardware support, particularly for building a computer that tripled the speed at which I could process my analyses using MrBayes. Finally, I would like to dedicate my thesis to my parents, Mrs Sheila English and Mr Michael English, for their constant love, support and encouragement throughout all my studies.

Chapter 1 Introduction

1.1 Overview

1.1.1 Central problems in HIV-1 vaccine research

1.1.1.1 The experimental and theoretical interface

The central problem for researchers of HIV-1 evolution is explaining the apparent design of the virus for causing pandemic infection in humans: understanding how HIV-1 spreads is key to halting the pandemic. The hypothesis of this thesis is that diversity of HIV-1 around the time of transmission is important for viral adaptation to a new human host, rather than the intrinsically superior transmissibility of particular strains circulating within human hosts, and studying recombination is fundamental for understanding transmissibility. Current knowledge of how HIV-1 spreads from host to host is based upon experimental observation and indirect inferences informed by theory. An experimentalist can draw upon current theory in bioinformatics for analysis of experimental data that would otherwise be intractable. Furthermore, qualitative theories of pathogen evolution can direct and focus experiments, as well as explain the experimental results. Lastly, robust and reproducible data acquired experimentally can be consistent with, or refute, current theoretical assumptions the mechanisms and dynamics of HIV-1 infection, and hence suggest new theoretical approaches.

1.1.1.2 Testing assumptions about HIV-1 transmission and within-host evolution

Models necessarily contain unrealistic simplifying assumptions. However, HIV-1 models have lacked scrutiny, and are accepted as

they reflect the limited available data collected from directly observing transmission of the virus. Data collection is challenged by technical and ethical limitations because HIV-1 produces an incurable infection, which results in social stigmatization or death. I address certain assumptions about HIV-1 transmission and evolution by investigating contemporaneous, multiple-variant transmission from a single host to two recipients and multiple-infection of a single host with epidemiologically unlinked strains.

1.1.2 The apparent design of HIV-1 for transmission

Transmission is the event that underpins natural selection because transmission determines the onward transfer of HIV-1 genetic material to new hosts. Darwinian evolutionary theory predicts that the apparent design of HIV-1 for human-to-human transmission results from the process of evolution by 'natural selection', where natural selection explains the heritable portion of systematic adaptation of HIV-1 to its ecological conditions (Darwin 1859; Frank 1996). In Darwinian evolutionary theory, the most common transmission scenario shapes adaptation by natural selection over many generations of infection (Frank 1994; Zuckerkandl and Pauling 1962). The most common scenario is sexual transmission of HIV-1 by a chronically infected host because most untreated hosts remain infectious for several years. Natural selection preserves 'favourable' traits and eliminates 'injurious' traits (Darwin 1859) that affect the probability of transmission (Fisher 1922). In the case of HIV-1, natural selection has shaped HIV-1 into a successful pandemic pathogen that has maintained an exponential growth rate for over thirty years (Barre-Sinoussi et al. 1983; UNAIDS 2010), underscoring its successful adaptation for transmission between humans.

1.1.3 Non adaptive variation, sampling error and bias

1.1.3.1 Challenges to inferring adaptive changes

It is challenging for experimentalists to identify adaptive changes that affect the probability of transmission (i.e. natural selection), given the presence of many other evolutionary forces and errors or biases in experimental data collection. The effect of natural selection represents only a fraction of the total evolutionary change that occurs through time (Fisher 1930). Non-selective evolutionary forces, which include any factor that creates phenotypic change through time but is not natural selection, also create genetic variation. These forces include random environmental effects, stochastic mutation and 'genetic drift' (Crow and Kimura 1965; Kimura 1968, 1969; Wright 1948). Genetic drift refers to the evolutionary change that occurs in a rapidly mutating population of limited size that may produce the appearance of systematic genetic variation (Crow and Kimura 1965; Kimura and Crow 1964; Lemey et al. 2009). However, the change is not adaptive because it is not generated by the process of natural selection (Kimura 1968, 1969; Wright 1948).

1.1.3.2 Challenges to identifying adaptive changes caused by experimental error

Experimental datasets of limited size or scope may be subject to significant errors or bias (Jordan et al. 2010). For example, a limited amount of genetic material is experimentally sampled, amplified and sequenced from the viruses present in the peripheral blood at a moment in time (Jordan et al. 2010). Sampling error for a host or population of hosts occurs both with respect to the total diversity present at any one moment as well as with respect to changes in diversity over time (Bull et al. 2009). Experimental bias may be introduced by repeated sampling of only one bodily compartment (Brown et al. 2011; Bull et al. 2009; Diem et al. 2008; Gantt et al. 2010; Heath et al. 2010; B. T. Korber et al. 1994; Lamers et al. 2009; Salemi et al. 2009) or by the laboratory methods used for sampling, amplification and sequencing (Jordan et al. 2010).

1.1.4 The effective population size

The most evolutionarily important genetic variants of HIV-1 are part of the 'effective population size', which represent the number of viruses that are successfully transmitted between humans (Wright 1938). The incidence of HIV-1 transmission per sexual act is generally low (Attia et al. 2009; Baggaley et al. 2010a, 2010b; Pilcher et al. 2004; Wawer et al. 2005), but the cumulative incidence of transmission averaged per infected person is sufficient to produce a pandemic (UNAIDS 2010). Over the course of chronic infection, substantial genetic variation is generated within a single host (Delwart et al. 1994; Shankarappa et al. 1999; Willey et al. 1986). However, by comparison the effective population size of transmitted viruses in infectious bodily fluids is tiny (Bull et al. 2009; Diem et al.

2008; Kemal et al. 2003). Furthermore, a severe genetic bottleneck is observed following transmission, particularly after recent sexual transmission (Derdeyn and Hunter 2008; Edwards et al. 2006; L. Q. Zhang et al. 1993).

1.1.5 Hard selection at transmission

Previous authors have hypothesized that the specificity of adaptation to conditions in the chronically infected host leads to 'hard selection' of genetic variation at the mucosal barrier of the new recipient host (Abrahams et al. 2009; Haaland et al. 2009; Haase 2005; Keele et al. 2008; Lee et al. 2009; Li et al. 2010; Pope and Haase 2003; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009). Hard selection, or mortality-driven selection, is caused by the instantaneous mass-attrition of genetic variants from the population of viruses contained in infectious bodily fluids before infection can be established (Haase 2005; Keele et al. 2009a). Hard selection contrasts with 'soft selection', which describes the selective outgrowth of one or more variants relative to one or more other variants. 'Soft selection' has been put forward as an explanation for the genetic bottleneck after HIV-1 transmission (Derdeyn et al. 2004; Kearney et al. 2009). However, supporters of a hard selection model also hypothesize that the genetic bottleneck may be caused by selection for rare adaptive qualities that facilitate or promote infection of a new host during the first hours or days following transmission, so that relative replication capacity (Keele et al. 2008; Ochsenbauer et al. 2012). In other words, they suggest that rare adaptive characteristics may be elucidated by inferring the exact variant that initiated the infection (Haase 2005; Keele et al. 2008; Ochsenbauer et al. 2012; Pope and Haase 2003). Furthermore, since

adaptive characteristics are apparently rare among viral variants within a host, these authors imply that the genetic bottleneck may be an adaptation by the virus to limit its initial diversity rather than the effect of an environmental constraint imposed by a newly infected host.

1.1.6 Acute infection and infectiousness

To identify potentially adaptive characteristics that enhance transmission, previous authors have focussed on the acute phase HIV-1 infections in newly infected hosts (Keele et al. 2008). Acute infection is characterized by the highest rate of onward transmission per sexual act as well as an exponential viral growth phase in the absence of an effective immune response (Pilcher et al. 2004). For example, the viral load peaks at approximately 10^7 virions per ml over a time scale of days to weeks, before the detection of neutralizing antibodies or strongly reactive cytotoxic T lymphocytes (Fiebig et al. 2003; Fiebig et al. 2005; Keele et al. 2008; R. E. Phillips et al. 1991). Viruses isolated during and shortly after the exponential viral growth phase spread rapidly between cultured cells *in vitro* (Dalmau et al. 2012). The acute exponential growth phase is co-incident with a precipitous depletion of CD4+ target cells in the peripheral blood, and the human immune response to HIV-1 is generally too late and ineffective to prevent irreversible damage caused by exponential virus growth during the acute phase (Clark and Shaw 1993).

1.1.7 Problems with inferring single variant infection

Previous authors have justified the single variant transmission theory because it provides a theoretically parsimonious explanation for the apparent genetic bottleneck in the face of limited

experimental data (Keele et al. 2008; Lee et al. 2009). However, the single variant theory involves strong assumptions that are based on indirect inferences rather than direct observations of the mechanics of transmission and evolution. For example, a key assumption is hard selection at the mucosal barrier (Keele et al. 2008; Lee et al. 2009; Pope and Haase 2003). There is little corroborating evidence to support a mechanism for hard selection at the mucosal barrier in most susceptible hosts. Furthermore, if hard selection appears to be for the collective good of the viral population, it is unclear how such an adaptation could be favoured by natural selection (Frank 1996). Lastly, certain assumptions in the single variant model lead to circular arguments. For example, single variant infections are analytically inferred in homogeneous datasets using a model fitted to observed infections that happen to be relatively homogeneous at the time of sampling (Keele et al. 2008; Lee et al. 2009).

1.1.8 Alternatives to single variant transmission

1.1.8.1 Compartmentalization

Another parsimonious explanation for at least part of the genetic bottleneck, corroborated by experimental observation, is that viral diversity within infectious genital fluids is much lower than total diversity, even in chronically infected donors of both sexes (Bull et al. 2009; Diem et al. 2008; Kemal et al. 2003). Compartment-specific haplotypes are likely to be only a transient phenomenon, due to rapid turnover, inter-compartmental migration and recombination (Brown et al. 2011; Bull et al. 2009; Diem et al. 2008; Kemal et al. 2003; Lamers et al. 2009; Philpott et al. 2005), and transient compartmentalization is also consistent with stochastic transmission of particular haplotypes.

1.1.8.2 Homologous recombination

Homologous recombination is explicitly ignored in the mathematical models for single variant transmission in HIV-1, despite the models containing the assumption of a neutrally evolving, panmictic population (Herbeck et al. 2011; Keele et al. 2008; Lee et al. 2009). If a particular variant grew at a faster rate than all other variants in the absence of diversifying selection by an effective immune response, then the viral population would come to resemble the dominant variant because of recombination. Even in a neutrally evolving, panmictic population, high rates of homologous recombination would be expected to have a homogenizing effect (Arenas and Posada 2010; Worobey and Holmes 2001). Unusually for an RNA virus, homologous recombination in HIV-1 has been measured *in vitro* to occur at a rate approaching that of nucleotide substitution (Hu and Temin 1990b, 1990a, 1992; Jones et al. 1993; Nikolaitchik et al. 2011; Temin 1993; J. Zhang and Temin 1994; Zhuang et al. 2002). Particular RNA motifs may increase the frequency of recombination (Klarmann et al. 1993), particularly between highly related sequences (Jones et al. 1993; Nikolaitchik et al. 2011; Zhuang et al. 2002). Recombination is frequently ignored in mechanistic models of within-host viral evolution *in vivo* because it is difficult to infer recombination among homogeneous sequences with any statistical confidence (Keele et al. 2008; Lee et al. 2009; Lemey et al. 2009). However, the justification for ignoring the contribution of recombination after transmission is another circular argument: recombination is difficult or impossible to detect in a homogeneous population so its potential to homogenize a viral population is ignored (Herbeck et al. 2011; Keele et al. 2008; Lee et

al. 2009). However, that recombination is difficult to detect does not mean that it does not occur or that its effect on observed evolution is insignificant.

1.2 Theoretical concepts underpinning the experimental approach

1.2.1 Dynamical models

Dynamical models in classical population genetics borrow concepts from physics and economics to make simple predictions of genetic change (Fisher 1930; S. P. Otto and Day 2007). In classical population genetics the evolutionary quantity of interest is the rate of change in the genetic frequency of a variant, which is measured as the percentage of a population. If the change can be shown to be due to natural selection, then the change in frequency is a measure of the variant's relative fitness. However, if the variant cannot be shown to be adaptive, then change in the relative frequency of the variant may be attributable to genetic drift or another non-selective evolutionary force.

1.2.2 The effect of diploidy on HIV-1 evolution

If one assumes that HIV-1 behaves as a diploid organism under natural selection, then otherwise inexplicable behaviour become logical. Classical models in population genetics begin with the assumption of haploidy or diploidy, meaning that each organism has one genomic copy or two genomic copies, respectively. The number of copies of the genome each organism possesses, termed the 'ploidy', affects predictions of genetic variant frequency change over time (Leigh 1990; S. P. Otto and Day 2007). Diploidy may be maintained by natural selection if recombination is beneficial (Barton and Charlesworth 1998; Haig and Grafen 1991). HIV-1 is a retrovirus, and is thus a member of the *Retroviridae* family of RNA viruses. Characteristic features that distinguish *Retroviridae* from

other RNA virus families are the occurrence of an obligate diploid (van der Kuyl and Cornelissen 2007) phase in the virus life cycle and relatively high recombination rates (Holmes 2009). The contribution of diploidy to the evolutionary process has traditionally been modelled in terms of Mendelian inheritance, with the simple null hypothesis being the Hardy-Weinberg equilibrium (Hardy 1908; Leigh 1990; S. P. Otto and Day 2007). Non-Mendelian organisms, such as RNA viruses, are generally modelled as haploid because most RNA viruses do not have an obligate diploid phase (Holmes 2009). As HIV-1 is a non-Mendelian diploid virus (van der Kuyl and Cornelissen 2007), it is sometimes referred to as pseudodiploid (W. S. Hu and Temin 1990a). The obligate diploid phase of HIV-1 occurs in the extracellular virion, which contains two copies of the RNA genome. The two RNA copies undergo reverse transcription into a single, haploid provirus prior to replication (W. S. Hu and Temin 1990a), rather than segregation into haploid gametes as in Mendelian organisms (S. P. Otto and Day 2007). However there are no explicit models for the effect of ploidy on predictions of pseudodiploid evolution that are equivalent to Mendelian diploid models.

1.2.3 Quantifying natural selection

1.2.3.1 Predictions of the change in the frequency of adaptive variant traits

Classical models in evolutionary biology apply a statistical approach to quantifying natural selection in terms of genetic trait frequency changes (Leigh 1990; S. P. Otto and Day 2007). Often the underlying mechanism for the genetic transmission of a trait is unknown. To determine whether a trait is adaptive or not, quantitative predictions

of trait frequencies in a population over a number of generations assuming that all change is caused by natural selection (Fisher 1930; Leigh 1990; S. P. Otto and Day 2007). Associating the frequency of a heritable trait with natural selection can also allow for initial qualitative predictions of changes in the frequency of a variant trait over time when the parameters of the system and the true mechanism of transmission are either unknown or not known with any certainty (Mao-Jones et al. 2010; Svara and Rankin 2011).

1.2.3.2 Short-term frequency change may not reflect long-term fitness

Short-term changes in the frequency of a genetic variant may not reflect the long-term relative fitness of that variant because there is a conflict between what is good for an individual and what is good for its group. For example, it is in the interest of a viral variant to try and transmit itself to a new host but, in doing so, it may reduce the chance of other viral variants being transmitted by killing the host. For example, a highly virulent strain may out-compete other variants for host resources but a race to use up host resources kills the host in a scenario known as 'tragedy of the commons' (Frank 1996; Hardin 1968; Kerr et al. 2006). In some cases, natural selection can favour the evolution of self-restraint, which may lead to a population that cooperates to preserve a chronically infected host. However, in the absence of any adaptations to maintain self-restraint, such a 'cooperative commons' would be unstable, as it may be invaded by a cheat (Frank 1996). Thus, pathogens that appear to optimize the probability of transmission from a chronically-infected host would be expected to evolve to limit virulence and prevent invasion of the population by cheats (Frank 1996).

1.2.3.3 Hard versus soft selection

Hard selection may be distinguished from soft selection primarily by the time scale over which it occurs. Hard selection occurs if variants are selected against by the complete death of a particular genetic mutant before reproduction can occur (Haldane 1932). The genetic history of the lineage that is selected against thus comes to an abrupt halt due to some absolute biological constraint. Soft selection, also termed 'slow' selection (Haldane 1932), occurs through differences in reproductive success between different genetic variants over a number of generations, rather than due to an absolute biological constraint.

1.2.3.4 Soft selection acts upon interacting individuals

Interactions between members of a population lead to evolution for social behaviours that complicate both the interpretation and quantification of heritable adaptation (Frank 1994). Non-sentient organisms, such as viruses, can evolve social behaviours that modify the probability of transmission (Frank 1994). For example, a population of viruses that undergo limited dispersal may not initially be in competition if surrounded by close relatives (Frank 1994). However, if relatedness decreases and the density of susceptible target cells falls below a critical threshold, the evolutionary balance may shift so that viruses evolve strongly competitive traits to ensure survival and transmission to a new host (Frank 1994, 1996). A shift to a more competitive phenotype provides a parsimonious interpretation for observations of HIV-1 during end-stage disease, when the viral load increases despite diminishing target cells and only limited evidence of relaxation in the host immune response (J. V. Giorgi et al. 1999; Huang et al.

2011).

1.2.4 Stochastic versus adaptive changes

1.2.4.1 Frequency-dependent stochastic change

In a frequency-dependent model, the chance of a strongly beneficial variant becoming extinct is dependent only upon its starting frequency within the population (Fisher 1922). As the frequency of the strongly beneficial mutant increases within the population, the chance of stochastic extinction decreases (Fisher 1922). The frequency-dependence assumption permits the hypothesis that the frequency of a variant will not change relative to its expected frequency unless natural selection has occurred (Fisher 1922, 1930; S. P. Otto and Day 2007). For example, the Shannon entropy score is used to screen for potentially adaptive mutations under the null hypothesis of no selection in consensus viral sequences sampled from large populations of infected hosts (Allen et al. 2005; B. T. Korber et al. 1994). However, the Shannon entropy score and similar methods are prone to being confounded by non-random sampling of non-adaptive genetic drift, called the 'founder effect' (Bhattacharya et al. 2007).

1.2.4.2 Genetic drift

If the chance of successful reproduction is very low for an individual HIV-1 variant within a large heterogeneous population, then the stochastic contribution to evolution will be large (Bhattacharya et al. 2007; Kimura and Crow 1964; Kimura 1968; Wright 1948). In a model where stochastic effects dominate (i.e. the effect of natural selection is assumed to be negligible), evolution can be represented as a process of 'genetic drift' through a continuous 'fitness landscape' (Wright 1948). In Mendelian terms, evolutionary

processes that act upon a haploid gamete are measured in terms of the consequences for the fitness landscape of the entire population of haploid gametes (Wright 1948). Natural selection produces rare vertical peaks and valleys in the fitness landscape and horizontal distance represents genetic distance, most of which represents selectively neutral variation (Kimura 1968, 1969; Wright 1948). Genetic distance is often simplified analytically to consider only mutation substitutions at homologous positions (Kimura 1968, 1969). Recombination allows for more efficient movement through the fitness landscape (Crow and Kimura 1965; S. P. Otto and Day 2007; Rambaut et al. 2004; Wright 1948) but is often ignored to simplify models for analytical tractability. In simplified models, evolution is assumed to occur much less frequently by recombination than by substitution (Lemey et al. 2004; Lemey et al. 2009; Zuckerkandl and Pauling 1962).

1.2.5 Interpreting natural selection

1.2.5.1 Inclusive fitness theory and the contribution of an individual virus within a population of viruses

Inclusive fitness theory formalizes prior assumptions into testable hypotheses, regarding whether or not an observed viral behaviour is adaptive, to allow interpretation of observed phenomena not satisfactorily explained by classical population genetics (Frank 1996; Hamilton 1964; Wild et al. 2009). For example, classical population genetics does not explain the complex pattern of genetic inheritance in HIV-1, which involves complex interactions between pseudodiploidy, recombination and other processes that are not directly observable in natural infection. Even ‘proximate’ (mechanistic) models (West et al. 2011) in post-classical population

genetics generate predictions based on simplified models of replication and mortality within simplistically-defined viral populations, predictions which may or may not diverge from reality when put into clinical practice (Duerr et al. 2012; Qureshi et al. 2012; Stephenson et al. 2012). In many biological settings, it is difficult to determine in observational studies which observed genetic changes are due to natural selection and which are due to genetic drift without some specific prior knowledge (Abbot et al. 2011; West et al. 2011). Prior knowledge is formalized by ultimate (functional) models, which rest upon the axiom that 'inclusive fitness' appears to be maximized by natural selection (Abbot et al. 2011; Dugatkin 2007; Frank 1996; West et al. 2011). Relatedness is a statistical measure to describe how similar the 'actor' and 'recipient' genotypes are relative to the population average (Hamilton 1964). The unit of selection, which is the level at which traits and relatedness values are measured, can be a whole virus or part of a virus genome in the context of rapid homologous recombination (Gardner and Grafen 2009). Inclusive fitness theory can generate qualitative or quantitative predictions of fitness (Abbot et al. 2011; Frank 1994, 1996; Hamilton 1964; West et al. 2011). The quantitative effect of inclusive fitness is the average effect that a single individual distributes among an entire population of recipients (Abbot et al. 2011; Hamilton 1964).

1.2.5.2 Relevance of inclusive fitness theory to HIV-1

Inclusive fitness theory is of relevance to HIV-1 for two main reasons. Firstly, the rapid turnover and high recombination rate of HIV-1 means that an individual virus is an unstable unit of selection (see 1.6). Secondly, classical population genetic models assume

Mendelian inheritance but, by assuming an average effect, one may adapt classical diploid models to pseudodiploid organisms such as HIV-1 (see Chapter 6).

1.2.6 Quantifying relatedness by common descent

1.2.6.1 Molecular Clock

A molecular clock quantifies relatedness in terms of the rate of mutation by substitution away from a single common ancestor (Gojobori et al. 1990; Kimura 1968, 1969; Zuckerkandl and Pauling 1962, 1965a, 1965b). Purely clock-like behaviour occurs when a population of constant size undergoes neutral evolution, with a constant intergenerational interval, from a single common ancestor (Holmes 2009; Lemey et al. 2009). If genetic variation in a neutrally evolving population of constant size demonstrates clock-like behaviour, one may retrospectively assume that the individuals of a population are kin, which means related by common descent (Lemey et al. 2009). Genetic sequences that do not demonstrate molecular clock-like behaviour may have experienced non-neutral evolution, recombination, extension, truncation, uneven intergenerational intervals or extreme changes in population size. In reality, purely neutral evolution in HIV-1 is rare (Holmes 2009; Lemey et al. 2006) and genetic bottlenecks are common between-hosts (Abrahams et al. 2009; Bar et al. 2010; Keele et al. 2008; Lee et al. 2009; Li et al. 2010; Ochsenbauer et al. 2012; Pope and Haase 2003; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009; L. Q. Zhang et al. 1993). Limited corrections can be made for non-clock-like behaviour if evolution is assumed to approximate neutrality with constant intergenerational intervals (Drummond et al. 2006; Holmes 2009; Keele et al. 2008; Lemey et al. 2006; Lemey et al. 2009).

Alternatively, assumptions of neutral evolution and kinship by common descent may also be used as null hypotheses for measuring the effects of natural selection or epistasis, where the phenotype of a gene or mutation may be altered by the effect of other genes or mutations (Edo-Matas et al. 2011; Shapiro et al. 2006). 1.2.6.2 Corroborating evidence for molecular clock estimates However, corroborating evidence is required for verifying retrospective estimates of the common ancestor derived from models of the molecular clock, due to the level of uncertainty in the estimates (Bos et al. 2011; Katzourakis et al. 2009; Pillay et al. 2007). Collecting corroborating evidence in rapidly evolving pathogens can be very challenging due to the extreme lack of preserved ancestral specimens (B. Korber et al. 2000; Worobey et al. 2008; Zhu et al. 1998) as well as uncertainty in reconstructing the direction of virus transmission between hosts (Pillay et al. 2007; Rachinger et al. 2011a).

1.3 Transmission

1.3.1 Overview

1.3.1.1 Determining the role of natural selection during transmission is potentially important for designing vaccination strategies

The key question for vaccine design is whether or not natural selection can be manipulated to prevent infections at the population level (Pope and Haase 2003). The adaptive fitness of a pathogen is defined as its ability to sustain transmission between hosts (Frank 1996). Darwinian evolutionary theory predicts that the within-host strategy of a virus is geared towards maximizing the chance of successful between-host transmission (Frank 1996). The process of HIV-1 transmission also involves a large stochastic element, since transmission success is rare among the diverse variants generated by a chronically infected host (Edwards et al. 2006; Rambaut et al. 2004; Wright 1948). If transmission is completely stochastic, then this represents a considerable challenge to designing an effective vaccine.

1.3.1.2 Key question for the study of HIV-1 transmission

Key questions remain when studying HIV-1 transmission:

1) What is the dominant infectious particle, cell-free virus or infected cells?

2) What characterizes the initial target cell or cells and how does the ability to infect this cell type or these cell types contribute to the genetic bottle-neck at transmission?

3) How does the infectious particle and/or the initial target cell influence subsequent viral evolution in a successfully infected host?

4) How can researchers realistically make predictions concerning the

above when experimental human infections are ethically impossible?
5) How can researchers extrapolate the results of experimental infections in non-human models to successfully design a preventative human vaccine?

Much of what is known concerning the biology of human transmission has been extrapolated from experimental animal studies or inferred from modelling of serendipitous sampling of human case studies. Human vaccination studies have also provided fresh insights but the lack of efficacy in trialled vaccines has also revealed critical gaps in the understanding of HIV-1 transmission. In this section, I will address how all these previous study types have yielded valuable information regarding the biology of HIV-1 transmission, but also how the data, analyses and interpretations remain flawed.

1.3.2 The biology of human transmission

1.3.2.1 Knowledge of the epidemiology of transmission

In epidemiological terms, transmission of HIV-1 from a donor host to a recipient host is well documented. HIV-1 is transmitted from a donor host to a recipient host following exposure to 'infectious' bodily fluids from the donor. Infectious bodily fluids contain donor host virus, and exposure to such fluids carries a relatively high epidemiological risk of acquiring infection. Infectious bodily fluids include blood, genital secretions (semen, pre-seminal fluid and vaginal fluid), rectal secretions and breast milk (Centers for Disease Control and Prevention 2012; J. A. Levy 2007; Zuckerman et al. 2004). By contrast other bodily fluids, such as saliva, are relatively non-infectious if uncontaminated by infectious fluids (Centers for Disease Control and Prevention 2012). Male latex condoms, which

prevent exposure to infectious bodily fluids during sexual intercourse, remain the most reliable methods of preventing transmission of HIV-1 during sexual intercourse (World Health Organization 2009). However, cultural, social, religious and economic barriers to condom use, as well as the risk of condom failure, necessitate the development of novel strategies for preventing HIV-1 transmission.

1.3.2.2 Uncertainty regarding biology of transmission in natural transmission

However, key aspects of the basic biology of transmission remain highly uncertain, hindering the development of novel prevention strategies. Uncertainty results from the ethical and technical barriers that currently prevent direct observation of natural human-to-human infection. For example, it remains unclear whether successful human-to-human transmission of HIV-1 is dependent upon exposure to free virus, infected cells, or both. Phylogenetic (T. Zhu et al. 1996) and animal (Salle et al. 2010) models of male-to-female transmission have previously indicated that both free donor virus and infected donor cells are capable of establishing infection in the donor host. However, even 'realistic' infections of experimental animals (Denton et al. 2011; Keele et al. 2009a; Letvin et al. 2007; B. L. Wei et al. 2005) employ cell-free fluid containing controlled concentrations of free virus (Letvin et al. 2007; B. L. Wei et al. 2005). If both free virus and infected cells can indeed seed infection, then prevention of infection by sterilizing immunity may require the induction of multiple types of immune response. Alternatively, topical vaginal gels containing a detergent (Van Damme et al. 2002) or the antiretroviral tenofovir (Abdool Karim et

al. 2011; Mayer et al. 2003; McCormack et al. 2010) have been trialled in at-risk women (Karim et al. 2011). Antiretrovirals should, in theory, prevent onward infection with sensitive virus regardless of the type of infectious particle. In practice, only topical antiretroviral gels have been partially successful, requiring high adherence to significantly impact on the heterosexual epidemic (Williams et al. 2011) and being impractical in cases of sexual assault. Furthermore, progress towards new preventative strategies is also hindered by poor characterization of the mechanisms by which HIV-1 breaches the physical and immunological barriers that protect the recipient host. Specifically, little is known about how HIV-1 traverses the mucous membranes involved during sexual transmission and, in particular, how transmission may be facilitated by mucosal inflammation or damage (Haaland et al. 2009). Topical anti-HIV-1 gels may induce or worsen existing mucosal inflammation and damage, paradoxically enhancing the risk of HIV-1 transmission (Van Damme et al. 2002). Even potential vaccines designed to prevent sexual transmission have been associated with futility in preventing, and possibly even enhanced risk of, HIV-1 acquisition (Buchbinder et al. 2008; National Institute of Allergy and Infectious Diseases 2007), indicating that there are critical gaps in current understanding of human-to-human transmission.

1.3.2.3 Initial target cells

Available data are conflicting as to the identity the initial target cells in vivo and the impact that this has on the subsequent host-pathogen interaction. Models based on rhesus macaques experimentally infected with the laboratory strain $SIV_{mac}251$ have indicated that CD4+/CCR5+ dendritic cells within the lamina propria

may be the earliest targets beyond the outer mucosal barrier (J. Hu et al. 2000; Spira et al. 1996) in male-to-female transmission. Complicating matters, one study has shown that the initial target cell type varies depending on whether rhesus macaques are inoculated subcutaneously with cell-free $SIV_{mac}239$ or $SIV_{mac}239$-infected cells (Ignatius et al. 2002). Furthermore, massive depletion of subsets of cells with low levels of CCR5 expression in acute SIV infection (Mattapallil et al. 2005) suggests a more complicated picture of co-receptor usage *in vivo*. However, the potentially non-trivial importance of the initial target cell type is supported by dominance of most natural HIV-1 infections by CCR5 phenotype viruses soon after sexual transmission in both men and women (Abrahams et al. 2009; Choe et al. 1996; Deng et al. 1996; Keele et al. 2008; H. Li et al. 2010; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009). Viruses with the CCR5 co-receptor usage phenotype are human macrophage cell-line tropic *in vitro*, and the ability to exploit CCR5 is evolutionarily conserved across HIV and SIV (Reeves et al. 1999). Furthermore, human hosts who genetically lack a functional CCR5 co-receptor are more resistant to sexually transmitted HIV-1 (Dean et al. 1996; R. Liu et al. 1996). Furthermore, experimental infection of rhesus macaques with highly pathogenic HSIV strains, which typically use CXCR4 in early infection, produces a different clinical picture with more rapid CD4+ T cell depletion compared to SIV, which typically uses CCR5 in early infection (Nishimura et al. 2005). Early human infections with CXCR4-using HIV-1 strains are also associated with increased CD4+ T cell activation that may lead to more rapid depletion (Hamlyn et al. 2012). While primary human cell isolates may well demonstrate

more complex co-receptor properties than *in vitro* cell lines, the dominance of both CCR5 co-receptors on freshly isolated target cells (B. Lee et al. 1999) from health donors and CCR5-using viruses in most early human infections suggests that the initial target cell population is of evolutionary significance to the virus.

1.3.2.4 Phylogenetic inference of HIV-1 transmission

Phylogenetic inference is frequently used to infer the dynamics of human-to-human HIV-1 transmission. However, the underlying assumptions used to reconstruct within-host evolution are necessarily simplistic to allow computation. For example, previous authors have justified a theory of 'single variant' transmission because it provides a theoretically parsimonious explanation for the apparent genetic bottleneck in the face of limited experimental data (Keele et al. 2008; Lee et al. 2009). Furthermore, it allows for a simple, Poisson-based probability model for describing diversification of HIV-1 in acute infection (Abrahams et al. 2009; E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009; Ochsenbauer et al. 2012; Salazar-Gonzalez et al. 2008). However, the single variant theory involves strong assumptions that are based on indirect inferences rather than direct observations of the mechanics of transmission and evolution. For example, a key assumption is that one may extrapolate transmission of a single whole virus from the observation that most early infections involve a homogeneous env population (Haaland et al. 2009; Keele et al. 2008; Keele et al. 2009a; H. Y. Lee et al. 2009). However, there is little corroborating evidence to support this extrapolation since it is not know to what extent envelope is under non-neutral selection during acute infection. In fact, arguments in favour of the theory of single variant

transmission contain circular arguments. For example, single variant infections are analytically inferred in homogeneous datasets using a model fitted to data from previously other observed infections that also happened to be relatively homogeneous at the time of sampling (Keele et al. 2008; Lee et al. 2009). This analytical model of early infection is necessarily fitted to previous data in this way because there is otherwise insufficient phylogenetic information within early, homogeneous infections to reconstruct an evolutionary history (E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009), in contrast to serial sequence sampling over longer time-scales following transmission (Lemey et al. 2006). Key data used to fit the single-variant transmission model parameters come largely from a small cohort of human hosts who were serially sampled during acute infection (Stafford et al. 2000), but this cohort of rare individuals may be prone to sampling errors, such as inclusion bias. Additional information in support of the single-variant transmission methods comes from animal studies, although such studies contain many known and potential flaws that may also bias interpretation (see 1.3.2.3 and 1.3.3). However, a critical piece of missing information from the data used to construct the single variant transmission model is the infection stage of the donors who infected the subjects used and the diversity that seeded the infections. Donors are themselves most infectious during acute or early infection (Wawer et al. 2005), when their own viral population would be expected to be relatively homogeneous. See Chapter 3 (sections 3.3.3.5 and 3.5.4) for further discussion of the single variant transmission theory and the tools employed to infer single variant transmission.

1.3.2.5 Barriers to infection at the mucosa

Human epidermal epithelium is physically protected by the thick stratum corneum, which contains multiple layers of dead cells. By contrast, mucous membranes are more physically vulnerable (S. Anderson et al. 2006; Coplan et al. 1996) and line the surfaces of relevance to HIV-1 transmission, for example surfaces in the rectal and genital tracts. Due to physical vulnerabilities to infection, mucous membrane barriers possess innate (see 1.3.2.6) and adaptive (see 1.3.2.7) immune defence barriers. These defences may be undermined by local inflammation, for example due to other sexually transmitted diseases (Haaland et al. 2009; J. A. Levy 2007) microtrauma (S. Anderson et al. 2006; Biggs et al. 1998) and toxic effects of introduced substances, such as spermicides (Van Damme et al. 2002).

1.3.2.6 Innate immune barriers to HIV-1 infection

The innate immune response does not rely on memory from previous exposures. Rather, the innate immune response recognizes evolutionarily conserved molecular features common to classes of pathogen. A lymphocyte subset, natural killer (NK) cells, recognize intracellular pathogens and kill infected host cells as part of the innate immune response (Alter et al. 2007b; Alter et al. 2007a; Bonaparte and Barker 2004). Of particular interest are the observations that 1) NK cells are represented in the mucosal linings of the genital and gastrointestinal tracts of uninfected hosts (Cella et al. 2009; Mselle et al. 2007; Mselle et al. 2009) 2) and NK cell populations expand to impede viral replication through cell killing when adaptive mechanisms of viral control are inadequate (O'Connell et al. 2009). Specifically, expression of class I HLA in complex with viral peptides is downregulated soon after viral

infection of a host cell, mediated by HIV-1 accessory genes (Kerkau et al. 1992) to evade the adaptive immune response. Recognition of viral peptides by NK cells relies principally on a lack of activation of inhibitory pathways, rather than activation by HLA-peptide complexes (Harel-Bellan et al. 1986). Inhibitory ligands on NK cells frequently recognize 'self' by interacting with class I HLA molecules (G. B. Cohen et al. 1999; Colonna et al. 1993; Harel-Bellan et al. 1986) via killer cell immunoglobulin-like receptor (KIR) (Yawata et al. 2006). Downregulation of class I HLA by HIV-1 should make the host cell more vulnerable to killing by NK cells, although this vulnerability may depend on which types of HLA molecule are downregulated (G. B. Cohen et al. 1999; Colonna et al. 1993). Indeed, the expansion of NK cells *in vivo* during acute and early infection is inversely correlated with CTL activity (Alter et al. 2007a). Furthermore, pathogen-activated NK cells can express cytokines, which induces more efficient killing of pathogens by cells of the adaptive immune system as well as macrophages. However, NK cells express CD4 as well as co-receptors for HIV-1, and so are susceptible target cells for HIV-1 infection (Valentin and Pavlakis 2003). In particular, NK cells express relatively high levels of surface CCR5 (B. Lee et al. 1999). CCR5-using virus, or viruses that are predicted to use CCR5 dominate early HIV-1 infection (see 1.3.2.3). Although NK cells are potentially protective against HIV-1 infection (Ravet et al. 2007; Scott-Algara et al. 2003), probably by impeding viral turnover (Alter et al. 2007b; Alter et al. 2007a), recruitment of large numbers of activated NK cells to the mucosal barrier would also risk recruiting large numbers of target cells for infection (Fauci et al. 2005; Valentin and Pavlakis 2003). Indeed, mucosal NK cells are

depleted early in infection.

1.3.2.7 Adaptive immune barrier to HIV-1 infection.

The adaptive immune response has been extensively studied due to variation in patient susceptibility and rate of progression to AIDS (see 1.4.3.4). It has been assumed that studying the robustness of the adaptive immune response at delaying or preventing progression to AIDS may uncover novel strategies for inducing a protective adaptive immune response. Inducing a protective adaptive immune response is the cornerstone of immunization by vaccination strategies. In particular, there is considerable interest in rapidly mobilizing a protective mucosal immune response to arrest onward infection of recipient host cells. Studies in rhesus macaques had suggested 1) a protective effect from neutralizing antibody at the mucosa (Mascola et al. 2000) and 2) that encouraging the production of a specific sub-population of memory T cells against SIV through vaccination would be key to inducing a sterilizing cytotoxic response on mucosal exposure to experimental free virus (Hansen et al. 2009). However, attempts to safely induce an adaptive immune response to prevent infection in humans have been largely unsuccessful. Early unsuccessful vaccine trials focussed on inducing an adaptive humoral response (Fauci and Fischinger 1988; Mascola et al. 1996; Mascola et al. 2000), using a vaccine template strategy that had been successfully employed against other viruses (Robey et al. 1986), such as hepatitis B virus (Stevens et al. 1980; Szmuness et al. 1980). Specifically, HIV-1 negative chimpanzee or human subjects were inoculated with the surface envelope glycoprotein of the virus (Fauci and Fischinger 1988; Mascola et al. 1996; Merz 1987). Despite initial promise in chimpanzees challenged with homologous viral

strains, these studies failed to induce an adequate neutralizing antibody immune response in human subjects. Only one study has demonstrated a partially protective effect to a vaccine designed to induce adaptive humoral and cytotoxic adaptive immunity, but this effect disappeared once the course vaccine booster doses had been completed (Rerks-Ngarm et al. 2009). Another study, which specifically sought to induce a cytotoxic response, was associated with an increased susceptibility to infection and was halted early on safety grounds (Buchbinder et al. 2008; National Institute of Allergy and Infectious Diseases 2007). The clearest danger when priming a CD4 T cell response is the recruitment of large numbers of potential target cells to the site of viral exposure.

1.3.3 Non-human models used to infer key events in HIV-1 transmission

1.3.3.1 Necessity for non-human animal models due to limitations of in vitro *experiments*

While important insights in relation to cell-to-cell virus transmission may be gained from *in vitro* experimentation with human cell culture systems, the complexity of a human host cannot be captured by *in vitro* experiments, necessitating the use of animal models. Specifically, host-pathogen interactions, such as viral selection driven by innate and adaptive immune responses at transmission and during acute infection, are of particular interest when trying to design preventative strategies, including vaccines. Since experimental infection of humans is ethically unacceptable, experimental infection of non-human models has been studied. However, experimental infection of great apes is also becoming increasingly unacceptable ethically, especially since the applicability

of non-human host-pathogen interactions to human infection have been questioned. (J. A. Levy 2007)

1.3.3.2 A brief overview of animal models

A number of animal models are used to simulate HIV-1 infection. These models include primate and non-primate models, such as humanized mice. Non-primate models are generally used to simulate very specific aspects of the human immune response to HIV-1, particularly where ethical restrictions prevent relatively trivial experiments in non-human primates. Humanized mice have been used to study prevention of infection by microbicide studies (Denton et al. 2011). However, non-primate models are uncommon in transmission studies and will not be discussed further.

1.3.3.3 Human specificity of HIV-1 for infecting human target cells prevents direct simulation of HIV-1 in non-human primates

HIV-1 is specifically adapted for transmission between humans and cannot produce persistent infection in non-human primates, despite simian immunodeficiency virus (SIV) strains infecting non-human primates being the closest extant genetic relatives to HIV-1. Species-specific restriction of SIV/HIV-1 infection in primates has been attributed in large part to viral adaptation to species-specific antiretroviral factors within host cells. A major group of innate antiretroviral factors that demonstrate species-specificity belong to the TRIpartite-Motif-5 (TRIM5) family, specifically TRIM5α (Kirmaier et al. 2010; Kratovac et al. 2008; F. Zhang et al. 2008). TRIM5α is a cytoplasmic protein that inhibits retrovirus replication in the cytoplasm soon after host-cell entry (F. Zhang et al. 2008). Furthermore, HIV-1 appears to have evolved accessory proteins that specifically overcome other innate barriers to retroviral infection in

human cells. For example, the viral vif accessory protein inhibits protective induction of hypermutation by human ApolipoProtein B Enzyme-Catalytic (APOBEC) proteins, specifically the human orthologue of APOBEC3G (Lecossier et al. 2003; Sheehy et al. 2002; Sheehy et al. 2003). HIV-1 has been re-isolated from captive chimpanzees two years following experimental infection (Gajdusek et al. 1985; Nara et al. 1987), consistent with both the degree of genetic relatedness between humans and chimpanzees as well as the likely common ancestor for HIV-1 and SIV_{cpz} (see 1.6.3.2). For example, HIV-1 group M strains and certain strains of SIV_{cpz}Ptt share key accessory gene functions when human cell lines are infected *in vitro*, for example vpu-mediated human CD4 and tetherin down-regulation (Kobayashi et al. 2011; Petit et al. 2011; Sauter et al. 2009; S. J. Yang et al. 2011). Some of these accessory gene functions are not shared, for example, between HIV-1 group M and other HIV-1 groups or more distantly related strains of SIV_{cpz} (Petit et al. 2011; Sauter et al. 2009; S. J. Yang et al. 2011). However, there are now significant ethical barriers to experimentation on great apes, and the lack of a consistently reproducible AIDS-like illness in experimentally infected further limits their attractiveness as models (Gajdusek et al. 1985; Keele et al. 2006; Keele et al. 2009b; Nara et al. 1987; Nath et al. 2000; Wadman 2011).

1.3.3.4 Natural SIV infection in African primates

SIV strains found in a subspecies of wild chimpanzee (*Pan troglodytes troglodytes*), SIV_{cpz}Ptt, have been inferred to be the closest genetic relatives of M group (pandemic) HIV-1 (Bailes et al. 2003; Gao et al. 1999; B. Korber et al. 2000) but ethical considerations have made experimental SIV infection in great apes,

specifically chimpanzees, extremely rare (Wadman 2011). SIV_{cpz}Ptt has been reported to be pathogenic in a single case study (Etienne et al. 2011). However, more convincing evidence for a causal link between an AIDS-like illness and SIV_{cpz} infection has been demonstrated in a colony of a different subspecies of wild chimpanzee, *Pan troglodytes schweinfurthii* (Keele et al. 2009b), which has been extensively studied over a number of years in Gombe National Park, Tanzania (Keele et al. 2009b; Santiago et al. 2003a; Santiago et al. 2003b). In this study, SIV_{cpz}Pts infection in *P. t. schweinfurthii* was associated with higher age-corrected mortality, and female infection was additionally associated with reduced fertility and increased infant mortality (Keele et al. 2009b). Furthermore, CD4+ T-lymphocyte depletion was demonstrated in the small number of infected chimpanzees where post-mortem samples were available (Keele et al. 2009b). The SIV_{cpz}Pts strain represents a distinct lineage that has evolved in relative isolation from the SIV_{cpz}Ptt strain and HIV-1 (Gao et al. 1999). However, previous authors have hypothesized that the high degree of genetic similarity between humans and chimpanzees (Drummond et al. 2006), and the hunting of chimpanzees for bushmeat by humans in west Africa, indicates both a common host susceptibility to related strains of pathogenic SIV_{cpz}Ptt as well as a mechanism by which the virus may have repeatedly crossed over to humans as a zoonosis and eventually evolved the ability to transmit between humans (Bailes et al. 2003; Gao et al. 1999; B. Korber et al. 2000). SIV_{cpz}Ptt has itself been inferred to be a recombinant strain derived from that found in collared mangabeys (*Cercocebus torquatus torquatus*), SIV_{Ctt}, and greater spot-nosed monkeys (*Cercopithecus nictitans*), SIV_{Cn}

(Bailes et al. 2003; Georges-Courbot et al. 1998). However, neither SIV_{Ctt} nor SIV_{Cn} appear to be pathogenic following infection of their natural hosts, and the true pedigree of pathogenic strains of SIV_{cpz}Ptt or M group HIV-1 remains poorly characterized (Bailes et al. 2003; L. Chakrabarti et al. 1987; Desrosiers 1990; Etienne et al. 2011).

1.3.3.5 SHIV model in pig-tailed macaques

Pig-tailed macaques (*Macaca nemestrina*) are an Asian primate species that lack a TRIM5α orthologue capable of blocking HIV-1 infection (Brennan et al. 2007; Brennan et al. 2008). Pig-tailed macaques can only be transiently, not persistently, infected with HIV-1 (Agy et al. 1992) because the HIV-1 accessory vif protein is unable to evade innate antiretroviral responses, especially APOBEC orthologues (Humes and Overbaugh 2011). However, persistent infection has been achieved in pig-tailed macaque *in vivo* by substituting a pig-tailed macaque SIV *vif* gene into an HIV-1 backbone, to form an "SHIV" chimera (Humes and Overbaugh 2011; Thippeshappa et al. 2011). SIV vif proteins are not as easily included in SHIV virions as in native SIV virions, so the level of pig-tailed macaque APOBEC3G orthologue inhibition is predicted to be reduced but nevertheless sufficient to achieve persistent detectable SHIV infection (Thippeshappa et al. 2011). However, pig-tailed macaques generate a potent immune response to SHIV (Thippeshappa et al. 2011), and do not demonstrate an AIDS-like disease in late SHIV infection when only the *vif* gene was replaced. Another strain, $SHIV_{SF}$1642.P4, is capable of producing an AIDS-like syndrome (Polacino et al. 2008), even with opportunistic *Mycobacterium tuberculosis* infection (Engel et al. 2012), in pig-tailed macaques. However, potent and protective neutralizing responses

can be reliably elicited against SHIV when pig-tailed macaques are immunized (G. B. Yang et al. 2012) partly because HIV-1 env gp120 protein must be passaged sufficiently to bind CD4 efficiently in pig-tailed macaques, but this evolved form is more susceptible neutralization (Humes and Overbaugh 2011). Thus the pig-tailed macaque model is qualitatively different enough to limit realistic simulation of natural HIV-1 transmission.

1.3.3.6 Rhesus macaque models

The rhesus macaque is particularly useful because 1) the rhesus macaque TRIM5α orthologue appears to have only minimal restriction activity against the laboratory strains of SIV, especially SIV_{cpz}, that are in current experimental use (Fenizia et al. 2011; Kirmaier et al. 2010) and 2) there are fewer ethical restrictions on rhesus macaque experimentation (Nath et al. 2000). The rhesus macaque monkey (*Macaca mulatta*) from Asia, particularly Indian sub-species, have been used to simulate HIV-1 infection since a cohort of macaques were noted to develop an AIDS like illness following accidental infection with SIV in laboratories (Desrosiers 1990; Kestler et al. 1990; Letvin et al. 1985). Experimental infection of rhesus macaques allows for controlled, reproducible studies of the earliest events after simulated sexual (mucosal) infection in ways (Salle et al. 2010; Spira et al. 1996) that it would be unethical to study in human subjects. The main advantage of the rhesus macaque model is this reproducibility of the AIDS-like disease process but over a shorter time-scale than in human infection (Kestler et al. 1990; Letvin et al. 1985). Another advantage of studying rhesus macaques is the large amount of data on this primate species available from extensive study in other fields of

biomedical and zoological research (Desrosiers 1990; Letvin et al. 1985).

1.3.3.7 Problems with the SIV rhesus macaque model in terms of viral phenotype

Important qualitative differences impact upon the realism of simulated human transmission in rhesus macaques. Firstly, $SIV_{mac}251$, $SIV_{mac}239$ and other common laboratory strains used to experimentally infect rhesus macaque monkeys in transmission studies have been highly and artificially selected as a result of a high number of *in vivo* and *in vitro* passages to facilitate infection of *M. mulatta* (Hirsch 2004; Kirmaier et al. 2010). For example, significant genetic evolution has occurred in the laboratory strain $SIV_{mac}251$ since it was first isolated for study (Strickland et al. 2011). Rhesus macaques cannot be infected with HIV-1 and are difficult to infect with primary isolates of SIV due to innate antiretroviral defences (Kirmaier et al. 2010). The commonest laboratory strains, including $SIV_{mac}251$ and $SIV_{mac}239$, are most closely related to SIV from sooty mangabeys monkeys (SIV_{smm}), which is in turn highly related to HIV-2 rather than HIV-1 (L. A. Chakrabarti 2004; Hirsch 2004; Naidu et al. 1988). HIV-1 and $SIV_{mac}239$ have critical differences at both the genomic and phenotypic level, including the observation that $SIV_{mac}239$ virions possesses a higher number of envelope surface spikes that are critical for transmission and cell entry (P. Zhu et al. 2006). Furthermore, HIV-2, $SIV_{mac}251$ and related strains demonstrate a significantly higher facility for infecting CD4 receptor negative/CCR5 co-receptor positive cells *in vitro* when compared with HIV-1 (Reeves et al. 1999) and there are functional differences between SIV_{mac} and HIV-1 in how the viral env gp120 trimer binds to CD4 *in vitro* (Finzi et al. 2012). Since SIV (Haase 2005; Keele et al. 2009a) and HIV-1 (Abrahams et al. 2009; Bar et al. 2010; Edwards et

al. 2006; Keele et al. 2008; H. Li et al. 2010; Salazar-Gonzalez et al. 2008; Shankarappa et al. 1999; T. Zhu et al. 1993) appear to undergo a considerable but variable evolutionary bottleneck around the time of sexual transmission, systematic differences between natural HIV-1 phenotypes and experimental macaque strain phenotypes may have non-trivial consequences for evolutionary predictions of natural HIV-1 transmission.

1.3.3.8 Problems with rhesus macaque models in terms of viral inoculation

Another non-trivial difference between natural HIV-1 infection in humans and experimental infection of rhesus macaque lies in the artificial experimental conditions of macaque infections. Reliable simulation of sexual transmission of SIV to uninfected rhesus macaque monkeys across mucous membranes has frequently required a much higher number of infectious particles than are estimated to be present in infectious human bodily fluids (Keele et al. 2009a; Rybarczyk et al. 2004; Stone et al. 2010). These higher infectious particle concentrations have been required due to innate retroviral restriction (see 1.3.3.3) as well as poorly characterized adaptive immune responses in highly exposed rhesus macaques (Hansen et al. 2009). However, experimental strains of SIV, like HIV-2, have a greater ability to infect CD4 negative cells than HIV-1 and have also been passaged through rhesus macaque tissue many times since primary isolation (see 1.3.3.7), which over time may have subtly biased the model. Recent studies of the rhesus macaque model have attempted to simulate natural human infection by repeatedly exposing uninfected animals to fluid containing a quantity of infectious particles similar to that seen infectious human

fluids (Hansen et al. 2009; Keele et al. 2009a; Stone et al. 2010). A critical flaw when trying to trace the transmission of particular *env* variants (Keele et al. 2009a; Stone et al. 2010) or to infer a useful vaccination effect (Hansen et al. 2009) is the artificially limited breadth of *env* diversity represented in the experimental inoculum, of what may already be an artificially infectious viral strain (see 1.3.3.7). Artificial limitation of the inoculum may not reflect the breadth of *env* diversity found in the infectious bodily fluids of chronically infected human male (Boeras et al. 2011; Brown et al. 2011; Delwart et al. 1998; Diem et al. 2008; T. Zhu et al. 1996) and human female donor hosts (Boeras et al. 2011; Kemal et al. 2003). In addition, previous macaque studies have focussed only on the genetics of cell-free virus in the inocula (Keele et al. 2009a; Stone et al. 2010), when the subtle balance between infectious particles type and initial target cell type may be disturbed by the significantly higher ability of SIV_{mac} strains to infect CD4 negative cells relative to HIV-1 (see 1.3.3.7).

1.3.3.9 Summary

Studies involving non-human primates have provided significant insights as to how HIV-1 and highly related strains of SIV are transmitted and interact with hosts. However, given the ethical and biological limitations of simulating human infection with non-human models, it is necessary to find alternative ways of studying real human infections. One such way is to develop more powerful analysis tools for studying natural human infection and to apply these tool to interesting human case studies that inadvertently mimic experimental animal study design (see Chapter 3).

1.3.4 Risk factors for human-to-human transmission and viral

evolution

1.3.4.1 Quantifying the risk of acquisition.

The lifetime risk of a susceptible host acquiring HIV-1 depends upon cumulative exposure to infectious virus in bodily fluids (UNAIDS 2010). Certain risk factors result in prolonged or repeated exposure to smaller quantities of virus, whereas other risk factors result in fewer episodes of exposure to larger quantities of virus.

1.3.4.2 High risk groups for acquisition

Groups at high risk of HIV-1 infection are young heterosexual men and women in sub-Saharan Africa, young homosexual men in developed countries, sex workers, intravenous drug users, prison inmates, the children of infected mothers, and those at risk of iatrogenic transmission in the absence of precautions such as mass screening of blood products (J. A. Levy 2007; UNAIDS 2010). Health care workers are also at risk through occupational accidents (J. A. Levy 2007; UNAIDS 2010).

1.3.4.3 The impact of risk factors on evolution

The largest influence on viral adaptation to host conditions is the most common transmission scenario within a population, and the HIV-1 epidemic has predominantly been a heterosexual epidemic at the global level (UNAIDS 2010). There has been a minor degree of genetic distinction between epidemics in homosexual men in developed nations and the dominant heterosexual epidemic, but this appears to be only a very recent development in the evolutionary history of HIV-1, with the advent of international air travel (Hue et al. 2005; UNAIDS 2010). Injecting drug use and iatrogenic transmission are also probably relatively recent risk factors in the evolutionary history of HIV-1 and account for a relatively small

number of transmissions at the global level (UNAIDS 2010). Mother-to-child transmission is relatively common and has probably existed throughout the evolutionary history of the epidemic (Epstein et al. 1985; Prado et al. 2009; Scott et al. 1989; Wolinsky et al. 1992). However, children usually die before causing onward infection and so this represents an evolutionary dead-end for HIV-1, meaning it would not have a large impact upon the evolution of the virus (Prado et al. 2009; Scott et al. 1989; Wolinsky et al. 1992). Thus sexual transmission, predominantly among heterosexual couples would be expected to have the greatest effect upon HIV-1 evolution, so one would expect HIV-1 to be best adapted for this mode of transmission.

1.3.5 Defining adaptive qualities for transmission

The fundamental evolutionary problem is reconciling the ability of HIV-1 to cause a pandemic with its apparently inefficient transmission if measured per sexual act (Attia et al. 2009; Baggaley et al. 2010b, 2010a; R. H. Gray et al. 2007c; R. H. Gray et al. 2007d; Haase 2005; Lingappa et al. 2010). HIV-1 is a very successful, pandemic pathogen. It has maintained an exponential growth rate in the human population for at least 30 years (Barre-Sinoussi et al. 1983; Serwadda et al. 1985), and probably longer (Worobey et al. 2008; Zhu et al. 1998). By contrast, some other RNA viruses, such as influenza, cause seasonal epidemics in human hosts (Holmes 2009). The rate of transmission between hosts is more rapid for influenza than HIV-1. Both HIV-1 and influenza are successful human pathogens but epidemic influenza occurs on a different time scale to epidemic HIV-1 (Holmes 2009). The different time scales for HIV-1 and influenza epidemics may be reflected in, but not necessarily

caused by, differing time scales of natural disease and virus eradication (Frank 1996). Influenza virus produces an acute infection and is eradicated relatively quickly, either by the human immune response or by the death of the host. In the absence of pharmacotherapy, HIV-1 produces a chronic infection, usually lasting several years, which only terminates with the death of the host except in rare cases of long-term viral suppression. HIV-1 may still transmit from one host to several hosts, but over a period of several years (Hayman et al. 2001). Therefore, HIV-1 appears poorly adapted for transmission per sexual act but is adapted instead for transmission due to the longevity of the infectious host and the cumulative exposure of susceptible hosts.

1.3.6 Transmission bottlenecks

Genetic bottlenecks are very common among diploid organisms during reproduction or transmission (S. P. Otto and Day 2007). Some authors assume that homogeneity and infectiousness in acute HIV-1 infection are qualities of a single founder virus that happens by chance to be a uniquely infectious clone (Haase 2005; Keele et al. 2008; H. Y. Lee et al. 2009; Ochsenbauer et al. 2012; Painter et al. 2003; Pope and Haase 2003; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009; Yeh et al. 2010). These authors argue that the enormous population size caused by the acute growth phase coupled with a rapid mutation rate allows the monophyletic virus to adapt and evade the immune response (Keele et al. 2008; Tsibris et al. 2009). However, it has been demonstrated that the pre-exponential phase of infection can last for a period of at least 31 days (Fiebig et al. 2005). Therefore, post-transmission evolution means that the consensus virus detected in the exponential phase

may not reflect a single original infecting strain.

1.4 HIV-1 disease and evolution

1.4.1 Overview

A critical balance for a pathogen is that between causing disease in the host with successful transmission. The eventual stable equilibrium for any fatal infection is the death of a susceptible host (Murray 2002). However, if there is a trade off between the pathogen's transmission rate and host survival, the pathogen maximises its overall transmission by reducing its transmission rate as this preserves the host for longer (Frank 1996). The pathogenic effects of HIV-1 are a product of this trade off as it exploits long-term chronic infection in the human host for the purposes of increased transmission (Frank 1996; J. A. Levy 2007).

1.4.2 Pathogenicity of HIV-1

1.4.2.1 Defining pathogenicity

All strains of HIV-1 are pathogenic, meaning capable of producing disease in a susceptible host (Frank 1996; J. A. Levy 2007), the end-stage of which is the acquired immunodeficiency syndrome (AIDS). All isolated viruses capable of replication are pathogenic *in vitro* due to direct and indirect mechanisms of cell injury (Blankson et al. 2007; L. Gray et al. 2007b; J. A. Levy 2007; Miura et al. 2010). HIV-1 is pathogenic even when isolated from hosts able to suppress viral replication to levels that are undetectable by conventional methods (Blankson et al. 2007; Miura et al. 2009b).

1.4.2.2 HIV-1 virulence and virulence factors

A spectrum of virulence exists within HIV-1. Virulence is the degree of pathogenicity as measured by the rapidity or severity of disease within a host. Virulence and factors determining virulence can only

be studied if one controls for host susceptibility and vice versa. Since both host susceptibility and viral virulence are both poorly characterized at present, controlling for confounding factors is not straightforward. One can compare extreme special cases such as rapid progressors to AIDS and elite suppressors of the disease. However, it is unclear to what extent observations from such special cases are generalizable to the majority of host-pathogen interactions. In addition, since pharmacotherapy has become safer and more effective at preserving life, it is becoming ethically difficult to observe cases with more or less rapid progression to AIDS. (J. A. Levy 2007)

1.4.2.3 Mechanisms of pathogenicity

HIV-1 is capable of causing target cell death directly and indirectly by cytopathic mechanisms. HIV-1 directly kills cells through toxic accumulation of viral products, including DNA and proteins, within the cell. Indirectly, viral infection induces activation and apoptosis of uninfected CD4+ T cells (Clerici et al. 1996; Finkel et al. 1995) and leads to the killing of infected cells by cytotoxic T lymphocytes (CTLs) (Brumme et al. 2008b; Frahm et al. 2004; Haas et al. 1996; Karlsson et al. 2007; McMichael and Phillips 1997; Payne et al. 2009; Perelson et al. 1996; R. E. Phillips et al. 1991). *In vitro*, HIV-1 causes single cell lysis and T cell tropic strains lead to fusion of cells to form syncitia (J. A. Levy 2007; Pantaleo et al. 1991). However, *in vitro* mechanisms of cell death differ in terms of importance from *in vivo* mechanisms of cell death. For example, syncitia are rarely isolated from cell specimens infected *in vivo,* except infected brain tissue in AIDS encephalopathy, compared with T cell tropic strains *in vitro* (Finkel et al. 1995; Sharer et al. 1985). It is also likely that the viral

nef gene, which encodes the nef protein and overlaps with transcription motifs, is somehow important for moderating pathogenicity (Deacon et al. 1995). Rare virus populations without fully functional nef proteins are associated with slower progression to AIDS (Deacon et al. 1995; L. Gray et al. 2007b; Kirchhoff et al. 1995). Nef-mediated down-regulation of host-cell CD4 may be related to enhanced pathogenicity (Aiken et al. 1994; Craig et al. 1998; Mariani and Skowronski 1993). However, Nef is promiscuous with its intracellular interactions, so its precise role in enhancing pathogenicity is poorly understood and may be complex (J. A. Levy 2007; Pizzato et al. 2007).

1.4.2.4 Observed virulence in early HIV-1 infection

The virulence of a pathogen may be influenced by competition for the chance to transmit to a new host (Frank 1996). In HIV-1, this model has been applied to the population collapse that follows the exponential growth phase of acute infection (A. N. Phillips 1996). In this way, the post acute phase collapse in the viral population can be explained parsimoniously without the necessity of a host immune response (A. N. Phillips 1996). When a pathogen population is growing in a host with finite resources, the within-host reproductive ratio will tend to decrease due to a finite 'carrying capacity' (S. P. Otto and Day 2007; A. N. Phillips 1996). Above the carrying capacity, the population size is unsustainable. If the human body is unable to regenerate target cells as fast as they are killed by HIV-1, then host resources are no longer constant and the carrying capacity will fall (A. N. Phillips 1996). Therefore, the fall in carrying capacity precipitates the observed collapse in the dominant, homogeneous viral population (Murray 2002; S. P. Otto and Day 2007). As the viral

load is collapsing, a rise in the concentration of cytotoxic T lymphocytes (CTLs) is observed in the peripheral blood (R. E. Phillips et al. 1991).

1.4.3 Host susceptibility

1.4.3.1 Defining host susceptibility

For a pathogen to cause disease or death, it must encounter a susceptible host. If HIV-1 invades a host who is susceptible to AIDS, then AIDS will occur in the absence of viral suppression with antiretroviral therapy. If a pathogen invades a human host that is not susceptible to AIDS, then disease will not occur even without antiretroviral therapy (Altfeld et al. 2003; O. J. Cohen et al. 1997; Miura et al. 2008; Miura et al. 2009b). HIV-1 in a non-susceptible host remains pathogenic but the host-specific factors that prevent the virus from causing disease remain poorly characterized at present (Blankson et al. 2007).

1.4.3.2 Observing host susceptibility to AIDS

Susceptibility to AIDS differs widely among human hosts. The estimated mean time to AIDS diagnosis is approximately eight years from seroconversion, with a 95% confidence interval of five to eleven years (J. A. Levy 2007). Mortality from AIDS in susceptible persons is 100 per cent without antiretroviral therapy (Kaufmann et al. 2004; Volberding and Deeks 2010). Susceptibility to AIDS was initially defined by the time between seroconversion and AIDS (J. A. Levy 2007; UNAIDS 2010). However, with the advent of highly active antiretroviral therapy (HAART) and close CD4+ lymphocyte monitoring in developed countries, rapid progression has been defined by a higher CD4+ lymphocyte threshold. At present, the UK guidelines recommend initiating treatment at a CD4+ lymphocyte

threshold of 350 cells/mm^3 or less (Gazzard et al. 2008; Volberding and Deeks 2010). There are plans in some countries to initiate HAART at a CD4+ lymphocyte threshold of 500 cells/mm^3 or less (Kitahata et al. 2009). Therefore, observing the natural history of HIV-1 infection is becoming increasingly less common for ethical reasons, necessitating the development of surrogate tests for susceptibility.

1.4.3.3 Categories of susceptibility to AIDS

Progressors are defined as those who progress to AIDS or who meet criteria for initiating HAART within twelve years of seroconversion without antiretroviral therapy (J. A. Levy 2007; Pereyra et al. 2008). Approximately 99% of human hosts infected with HIV-1 are progressors (Grabar et al. 2009; J. A. Levy 2007). Therefore, the vast majority of infected humans are susceptible to disease and death caused by HIV-1. Rapid progressors are defined as those who progress to AIDS or who meet criteria for initiating ART within four years of seroconversion without antiretroviral therapy (Kaslow et al. 1990; J. A. Levy 2007). Approximately ten per cent of human hosts infected with HIV-1 are rapid progressors (J. A. Levy 2007). Long-term non-progressors account for less than one per cent of infected human hosts and are defined as those who have not progressed to AIDS, or who meet criteria for initiating HAART within four years of seroconversion without antiretroviral therapy (Grabar et al. 2009; J. A. Levy 2007). However, host-specific mechanisms by which long-term non-progressors remain disease-free for such long periods, or indefinitely, remain unknown.

1.4.3.4 Mechanisms restricting host susceptibility to AIDS

The quality and strength of the immune response are likely to be the

most common determinants of host susceptibility to AIDS, but other known mechanisms also exist and susceptibility may be multi-factorial (Cohen et al. 1997; McMichael and Phillips 1997). Certain mutations that affect viral replication *in vitro* lead to viral escape from immune responses associated with particular Class I human leukocyte antigen (HLA) alleles (Brockman et al. 2010; Brumme et al. 2011; Miura et al. 2009a; Miura et al. 2010) and viral load during chronic infection *in vivo* (Frahm et al. 2006; Frater et al. 2007; Goepfert et al. 2008; Kiepiela et al. 2004; Martinez-Picado et al. 2006; Matthews et al. 2008; R. E. Phillips et al. 1991). There is also evidence of viral adaptation to HLA-restricted immune responses at the human host population level (Matthews et al. 2009; Payne et al. 2009). Some immune escape variations revert to the consensus type upon transmission to a host lacking the relevant Class I HLA allele. However, most reversion to the consensus is slow (Duda et al. 2009; Fryer et al. 2010; Fryer and McLean 2011). CTLs can suppress viral replication by recognizing and killing HIV-1 infected cells *in vitro* that present HLA-matched epitopes (Klenerman et al. 1996; Payne et al. 2010). The quality and maturation of the cytotoxic immune response is also associated with viral control *in vivo* (Day et al. 2007; Rinaldo et al. 1995). Furthermore, infants and young children tend to progress more rapidly to AIDS than adults and this may reflect qualitative differences with the adult immune response (Fauci 1988; Scott et al. 1989). In addition, some people genetically lack a fully functional version of the common co-receptor CCR5 that is exploited by HIV-1 for cell entry (Allers et al. 2011; Benkirane et al. 1997; Y. Huang et al. 1996). Non-human hosts are generally not susceptible to HIV-1 infection because of intracellular antiviral mechanisms that

are lacking or have been subverted by HIV-1 in humans (Kratovac et al. 2008; Thippeshappa et al. 2011; F. Zhang et al. 2008). However, none of these known host factors are absolutely protective from progression to AIDS.

1.5 Multiple-infection

1.5.1 Overview and definitions

Multiple-infection is a major source of HIV-1 diversification and may play a critical role in viral adaptation to individual hosts as well as to populations of hosts (Koelsch et al. 2003; Rachinger et al. 2011a; D. M. Smith et al. 2005b; D. M. Smith et al. 2005a; Streeck et al. 2008; van der Kuyl and Cornelissen 2007). Multiple-infection is defined as

Table 1.1 Summary of key definitions

Term	Definition
Multiple-infection	Invasion of an epidemiologically linked population of viruses within a single host by a strain originating in an epidemiologically unlinked host so that both strains coexist.
Superinfection	Replacement of the original strain by an invading strain or a recombinant between the original and invading strains.
Co-infection	Contemporaneous transmission of two or more strains to a host during primary infection.

invasion of an epidemiologically linked population of viruses within a single host by a strain originating in an epidemiologically unlinked host so that both strains coexist. Multiple-infection may be stable or progress to the extinction of either strain, resulting in superinfection. The term ‘dual infection’ has been used by previous authors (van der Kuyl and Cornelissen 2007) but hosts may be multiply–infected by more than two epidemiologically unlinked strains (Takehisa et al. 1997; van der Kuyl et al. 2005). The term ‘superinfection’ is sometimes confused with multiple-infection but super-infection is the complete replacement of viral strains by a dominant viral strain. Instead of a clear replacement of all strains by an invading strain, the strain that eventually wins out may be a

recombinant species between the multiply infecting strains and so superinfection is may be used as surrogate evidence of multiple-infection (Altfeld et al. 2002; McCutchan et al. 2005; Streeck et al. 2008; van der Kuyl and Cornelissen 2007). 'Co-infection' is defined as contemporaneous transmission of more than one epidemiologically unlinked strain from a multiply infected host, but authors may assume co-infection if multiple-infection is detected at the earliest available blood sample (van der Kuyl and Cornelissen 2007).

1.5.2 Prevalence of multiple-infection

A necessary condition of multiple-infection is non-contemporaneous exposure to more than one epidemiologically unlinked strain (van der Kuyl and Cornelissen 2007). Detection of multiple-infection or recombinant superinfection is uncommon, even in highly exposed hosts (Rachinger et al. 2011b).

1.5.3 Factors which limit multiple-infection of a host

1.5.3.1 Relative fitness of viral variants

Superinfection may theoretically lead to increased virulence (Nowak and May 1994; van der Kuyl and Cornelissen 2007). Dual infections have been demonstrated in long-term non-progressors (Casado et al. 2007; Rachinger et al. 2011a), where the replication capacity of the original strain may be reduced by the cost of evading eradication by the immune system (Brumme et al. 2011; Miura et al. 2009b; van der Kuyl and Cornelissen 2007).

1.5.3.2 Cross-reactive immunity

There is evidence that some infected hosts generate broadly cross-reactive antibodies. However, superinfection may still occur even in the presence of an apparently robust and broad immune response

(Altfeld et al. 2002).

1.5.3.3 Depletion for target cells from mucosal surfaces

Since most HIV-1 infections occur by HIV-1 breaching a mucosal membrane, there has been considerable interest in the apparently selective depletion of target cells from these regions during primary infection (Brenchley et al. 2004; Guadalupe et al. 2003; Mehandru et al. 2004). A similar pattern of mucosal cell depletion is seen in SIV infections of non-human primates (Gordon et al. 2007; Q. Li et al. 2005b). However, depletion of mucosal target cells leaves fewer target cells that can be exploited by exogenous virions. Thus, early depletion of mucosal target cells may be an adaptation to limit multiple-infection or superinfection beyond the first few months following seroconversion.

1.6 Homologous recombination

1.6.1 Overview

Homologous recombination is largely a stochastic mechanism of mutation in that it occurs more commonly among highly related sequences. HIV-1 has an obligate diploid virion stage in its lifecycle. The evolutionarily effective rate of recombination is enhanced or constrained by a number of factors that remain poorly characterized. Recombination unlinks beneficial genetic variations from deleterious genetic variations and allows for the efficient evolution of the pathogen. Recombining viruses undergo evolution by genetic drift and natural selection. Recombination may facilitate diversification or homogenization of a viral population depending upon the balance of evolutionary forces in a system.

1.6.2 Mechanisms of homologous recombination

1.6.2.1 Key events in retroviral replication

Retroviruses have two copies of an RNA genome that are reverse transcribed into a single DNA copy by a viral enzyme called reverse transcriptase. The diploid virion enters the host cell through interactions between the virus envelope proteins, encoded by the viral *env* gene, and surface molecules on the host cell, including the CD4+ receptor as well as various co-receptor (Connor et al. 1997) and enhancer molecules. After cell entry, the diploid RNA genome is reverse transcribed into a haploid complementary DNA (cDNA) copy *en route* from the host cell membrane to the nucleus. The enzyme reverse transcriptase, encoded by the viral *pol* gene, is responsible for making the cDNA copy, and the enzyme is particularly prone to introducing point mutations and recombination during reverse

transcription. Before virion assembly, a single transfer RNA (tRNA) molecule is annealed to the diploid genome to act as a primer for reverse transcription on one of the strands. The reverse transcriptase enzyme switches between RNA strands during reverse transcription. The resulting single cDNA copy is turned into a haploid double-stranded DNA copy, called a provirus, by the same reverse transcriptase enzyme. The double-stranded provirus is then inserted into the genome of the host cell by another viral enzyme called integrase, also encoded by the viral *pol* gene. Following proviral integration, the host cell machinery manufactures both viral proteins and RNA genomes. The production of certain accessory viral proteins, such as *nef*, leads to down-regulation of CD4, limiting entry of other virions into the cell, although there is a very limited barrier to multiple-infection of cells *in vitro* (D. N. Levy et al. 2004). Diploid HIV-1 virions are assembled at the host cell membrane (Clever et al. 1995; Jones et al. 1993; Mark-Danieli et al. 2005; Sakuragi et al. 2003), forming the viral envelope from the lipid bilayer and membrane-associated molecules of the host cell as well as a limited number of viral envelope proteins. HIV-1 virions may be 'pseudotyped' with envelope proteins that have a different proviral template *in vitro* (Mascola et al. 2005) but the significance of this effect *in vivo* is unknown. HIV-1 also causes cell-to-cell fusion *in vitro* (Pang et al. 2000; Sato et al. 1992) but the contribution of this action to recombination *in vivo* remains unknown.

1.6.2.2 Impact of sequence divergence on co-localization of strains in a virion

The rate of recombination is partly limited by the ability of two strains to join together for packaging in a virion, so that more divergent strains are less likely to dimerize (Chin et al. 2005; Nikolaitchik et al. 2011). These two copies of the RNA genome are joined together within the host cell, in a process called dimerization. After the genomes dimerize, the structural proteins of the virion surround the genome, in a process called encapsidation, as part of virion assembly (Clever et al. 1995; Jones et al. 1993; Mark-Danieli et al. 2005). Dimerization of the two RNA genomes prevents non-genomic RNA and other molecules from being packaged within the virion (Clever et al. 1995). Dimerization occurs between conserved motifs in the secondary RNA structure of each genome (Nikolaitchik et al. 2011). The dimerization initiation signal (DIS) motif, located in the untranslated 5´ long terminal repeat (5´LTR) region of the genome, is particularly conserved (Chin et al. 2005; Damgaard et al. 2004). Critical mutations can significantly reduce dimerization affinity, in some cases by 100 fold (Chin et al. 2005; Lu et al. 2011; Nikolaitchik et al. 2011; Sakuragi et al. 2003). Encapsidation is also limited by critical mutations in the RNA structure that reduce binding affinity between the dimmer and the structural nucleocapsid protein, which is encoded by the viral *gag* gene (Lu et al. 2011).

1.6.2.3 Impact of sequence divergence on template switching during reverse transcription

Retroviral recombination occurs primarily when the reverse transcriptase enzyme switches between RNA strands when reverse transcribing the single DNA provirus (Hu and Temin 1992; Jones et al. 1993; Luo and Taylor 1990). Homologous recombination is at least 100 times more common than non-homologous recombination (J. Zhang and Temin 1994). Certain conserved motifs within the RNA secondary structure promote template switching at particular locations in the genome (Galetto et al. 2004; Moumen et al. 2001; Zhuang et al. 2002). Reverse transcriptase does not switch between RNA and cDNA templates during double-strand DNA provirus synthesis because the RNA strands are first degraded by RNase H (Huber et al. 1989). Rather, degraded RNA molecules may act as primers for synthesis of the second-strand, resulting in recombination by a less-common, alternative mechanism (Fuentes et al. 1996).

1.6.2.4 Defining the units of selection

A virus may not act as a single unit of selection: instead it may be better to consider genes or genomic regions as the unit of selection. As an individual virus exists only transiently, it is not a sufficient unitary entity for natural selection to act upon (Gardner and Grafen 2009). Therefore, each gene will appear as though it is trying to maximize its own fitness (Gardner and Grafen 2009; Hamilton 1964). For example, the gene for a neutralization-resistant surface protein may recombine with another gene that promotes rapid replication, and so to boost its chance of transmission.

1.6.3 Recombination and the evolutionary origins of HIV-1

1.6.3.1 HIV-1 sub-classification and recombination

HIV-1 has been subdivided into four groups based upon phylogenetic clustering. The most commonly detected phylogenetic cluster is the major group (M group) and M group causes pandemic HIV-1 infection. Pandemic M group has the broadest genetic diversity of all HIV- groups, including recombinant diversity. Group M is divided into subtypes, circulating recombinant forms and unique recombinant forms. M group subtypes and circulating recombinant forms reflect isolation of the virus into recognizable clusters of transmission (Kalish et al. 2004; Pybus et al. 2000; Robertson et al. 2000). Viral isolation is often geographical but may also reflect well-connected international networks of high-risk groups (Hue et al. 2005). Unique recombinant forms represent inter-subtype recombination that has not formed a transmission cluster (D. M. Smith et al. 2005a). At least ten per cent of circulating HIV-1 diversity involves inter-subtype recombination (Magiorkinis et al. 2003), although identification is limited by imputation of parental strains (J. P. Anderson et al. 2000) and it is highly probable that circulating ‘subtypes’ also have recombinant origins (Rambaut et al. 2004).

1.6.3.2 SIV in chimpanzees

It is hypothesized that the HIV-1 M group resulted when SIV crossed from chimpanzees to humans. As previously discussed (see 1.3.3), SIV strains exist in many non-human primates found in Africa (Bailes et al. 2003; Franchini et al. 1987; Fukasawa et al. 1988; F. Gao et al. 1999; Holmes 2009; Keele et al. 2009b; Thippeshappa et al. 2011). Phylogenetic evidence is consistent with a common ancestor for HIV-1 M group and SIV found in a subspecies of chimpanzee (*pan troglodytes troglodytes*) which now inhabits a restricted territory of west Africa (F. Gao et al. 1999; Holmes 2009; Keele et al. 2006; Rambaut et al. 2004). Furthermore, the central chimpanzee subspecies is the only primate of African origin that is known to be susceptible to AIDS apart from the human species (Keele et al. 2009b). The strain of chimpanzee SIV (SIV_{cpz}) found in *p. t. troglodytes* is hypothesized to be a recombinant between two divergent strains of SIV found in two different species of monkey (Bailes et al. 2003). The inferred parental viruses, based upon a limited range of sampled SIV sequences, are found in collared mangabeys (*Cercocebus torquatus torquatus*) and greater spot-nosed monkeys (*Cercopithecus nictitans*) (Bailes et al. 2003). The modern-day territories of collared mangabeys and greater spot-nosed monkeys overlap with that of *p. t. troglodytes* (Bailes et al. 2003). Consistent with the theory that SIV was able to cross into humans due to recombination, artificially recombinant HIV/SIV chimeras have been generated experimentally that are capable of causing persistent infection in African pig-tailed macaques (*Macaca nemestrina)* which are resistant to persistent HIV-1 infection (Thippeshappa et al. 2011).

1.6.4 Consideration of some theoretical disadvantages of recombination for HIV-1 adaptation within hosts

Recombination is partly a stochastic process and may not always appear to be selectively advantageous if only viewed at the population level. For example, recombination may not contribute to the "heritable variance of fitness" (Fisher 1930) if it disrupts links between favourable mutations as often as it disrupts links between favourable and unfavourable mutations (Barton and Charlesworth 1998). The loss of advantageous combinations of mutations, known as haplotypes, is the most obvious cost of recombination to the individual (Barton and Charlesworth 1998). However, a higher rate of recombination between highly related strains of HIV-1 might mitigate this effect if, for example, an advantageous gene is less likely to recombine with an injurious gene. Conversely, it has also been shown experimentally in other RNA viruses that recombination between co-circulating strains can have a negative impact on the ability of a viral population to purge deleterious mutations *in vitro* (Froissart et al. 2004). Some authors have argued that interpretations of recombination at the population level often involve implicit or explicit assumptions of group selection, which has been discredited as a theory (Feldman et al. 1996). However, disruption of linkage disequilibrium can have important benefits if one considers the unit of selection to be individual genes, rather than viruses (Gardner and Grafen 2009). For example, under the influence of strong selection *in vivo*, such as from the human immune system, recombination appears to have a beneficial diversifying effect upon pathogens (Guttman and Dykhuizen 1994). Furthermore, rescue from fatal mutations attributable to

recombination has been demonstrated *in vitro* (Clavel et al. 1989). If between-host transmission includes a strong stochastic element and severe bottleneck, then it may be that recombination is necessary for adaptation to new host if a limited degree of disruption to linkage equilibrium is advantageous to the individual in terms of inclusive fitness.

1.7 Summary of work

1.7.1 Hypothesis

The hypothesis of this thesis is that diversity of HIV-1 around the time of transmission is important for viral adaptation to a new human host, rather than intrinsic superiority of particular strains found in infectious fluids from human donor hosts, and studying recombination is important for understanding this behaviour.

1.7.2 Identification of key questions

1.7.2.1 Key questions

The two key questions addressed in this thesis are:

1. Is there any evidence of hard selection of viral diversity at transmission?

2. To what extent does recombination play a role in shaping viral diversity in multiple-infection, when two variants of the same subtype co-circulate at very different initial frequencies?

1.7.2.2 Summary of approach to answering key question 1.

My hypothesis is that transmission of HIV-1 *env* gene variant diversity from one human host to another appears to occur at random. However, the apparent randomness of *env* variant transmission does not imply that HIV-1 evolves at random around the time of transmission. Rather, HIV-1 is highly adapted for transmission among humans and once infection has been established, it cannot be eradicated by the human host. To demonstrate the apparent randomness of *env* transmission, I test the null-hypothesis that hard selection accounts for between-host viral divergence in a rare case study of contemporaneous infection. My approach contributes to the current literature on HIV-1,

specifically by proposing and investigating an alternative hypothesis to the single virion theory of sexual transmission.

1.7.2.3 Summary of approach to answering key question 2.

My hypothesis is that recombination is an important behaviour for successful HIV-1 transmission and adaptation to a new host. Instead of ignoring or discounting recombination, as is the case with most approaches to understanding HIV-1 evolution, I bring recombination to the forefront of my study of multiple infection. I demonstrate that multiple infection and recombination in the absence of selection pressure by the immune system can be readily explained if one assumes that HIV-1 behaves like a diploid organism under natural selection. My approach contributes to the current literature on HIV-1 by characterizing recombination following multiple infection and using the distinct properties of a diploid model, versus a haploid model, to explain observed viral sequence changes in the absence of an effective, detectable immune response in the host.

1.7.3 An outline of the contents of each chapter

In Chapter 2, I describe the methodologies used. In Chapter 3, I describe a special case of transmission in which I infer that a single donor infected two recipients contemporaneously with genetically divergent viruses. In addition, I demonstrate that this divergence is not attributable to detected immune responses in either patient, consistent with stochastic transmission, rather than hard selection, of diversity. In Chapter 4, I quantify the between-host diversity in four genes in terms of time to the most recent common ancestor. In Chapters 5 and 6, I present additional experimental data and a novel approach to inferring epidemiologically unlinked, intra-subtype multiple-infection in one of the recipients. Additionally, in Chapter 6

I infer persistent multiple-infection and recombination between epidemiologically unlinked strains before eventual superinfection by a recombinant strain, consistent with selection. In Chapter 7, I discuss the relevance of these findings and how they challenge some of the basic assumptions in the single variant transmission theory.

1.7.4 Summary of experimental data

Table 1.2 Summary of experimental data

Data	Description	Chapters
Cohorts		
SPARTAC (England)	Clinical trial	3, 4, 5, 6
St Mary's Hospital Acute Infection Cohort	Case control study	3, 5, 6
Thames Valley Cohort (Reading)	Observational study	3
Sequencing strategies		
Population PCR	-	3, 5
Bacterial cloning	-	3, 4, 6
Single genome amplification	-	3, 4, 6
Sanger sequencing	-	3, 4, 5, 6
Viral genes		
gag	-	3, 4, 5, 6
pol	-	3, 4
env	-	3, 4, 5, 6
nef	-	3, 4, 5
Gene fragments		
gp160 coding region to the end of the *gp120* coding region	(HXB2 nucleotide position 6225 to 7758)	3, 4, 5, 6
env C2V5	(HXB2 position 6810 to 7631)	3, 4
short *gag* (*p24* to	HXB2 1471 to	3

p6)	1976			
long *gag (p17* to *p7)*	HXB2 2042	790	to	3, 4, 5, 6
nef-containing (*gp41* to within the *3´LTR)*	HXB2 9532	8289	to	3
nef-only	HXB2 9417	8797	to	3, 4, 5
pol RT	HXB2 3428	2643	to	3, 4

Chapter 2 Materials and Methods

2.1 Overview

The projects described in this thesis used a combination of modelling, observed data and analysis to elucidate some fundamental aspects of HIV-1 evolution and its relevance to vaccine design. Chapter 2 outlines explanations of and justifications for experimental and analytical techniques used in later chapters.

2.2 Subjects

2.2.1 Definitions and abbreviations

See Table 2.1 for a list of terms and abbreviations.

Table 2.1 Abbreviations

Abbreviation	Term
Ab	Antibody
Ag	Antigen
ELISA	Enzyme-Linked ImmunoSorbent Assay
SPARTAC	Short Pulse AntiRetroviral Therapy at seroConversion
WB	Western Blot

2.2.2 Cohorts

2.2.2.1 Summary of cohorts

See Table 2.2 for a summary of the cohorts used in this thesis.

Table 2.2 Summary of cohorts

Cohort	Geographical location	Dates of recruitment	Number of participants
1. SPARTAC (England)	London, Birmingham, Brighton	2003 to 2007	148
2. St Mary's Hospital Acute Infection Cohort	London	1999 to 2003	105
3. Thames Valley Cohort (Reading)	Reading	2006 to present	300+

2.2.2.2 SPARTAC trial

360 participants were recruited to the Short Pulse AntiRetroviral Therapy at HIV seroConversion (SPARTAC) trial (ISRCTN number 76742797; EudraCT number 2004-000446-20). 148 participants were from England. Participants were at least 18 years of age. Intravenous drug use was a criterion for exclusion.

2.2.2.3 St Mary's Acute Infection Cohort

105 participants were recruited into the St Mary's Hospital Acute Infection Cohort. This cohort was considered a pilot study for SPARTAC (Fidler et al. 2002).

2.2.2.4 Reading transmission recipient pair

Two participants, R1 and R2 were consented and recruited from the genitourinary medicine clinic of the Royal Berkshire Hospital in Reading, UK, as part of the Thames Valley Cohort. Clinical information and specimens were kindly provided by Dr Fabian Chen, Professor Philip Goulder and Dr Philippa Matthews.

2.2.3 Fiebig Staging

2.2.3.1 Fiebig staging method

Fiebig staging (Table 2.3) characterizes the order of clinical test positivity in primary HIV-1 infection, based upon two rare datasets of archived, serial samples from blood donors (Fiebig et al. 2003; Fiebig et al. 2005).

Table 2.3 Summary of Fiebig stage definitions from the onset of detectable viraemia (day 0). *(Fiebig et al. 2003; Fiebig et al. 2005)*

Fiebig Stage	Laboratory definition
Pre-Fiebig	Viraemia <100 copies/ml or undetectable
I	Viraemia only (≥100 copies/ml)
II	Viraemia+ p24 Ag
III	Viraemia + ELISA +/- p24 Ag
IV	Viraemia + ELISA + indeterminate WB
V	Viraemia + ELISA positive + WB (but not p31 Ag)
VI	Viraemia + ELISA positive + WB (including p31 Ag)

The staging method also estimates the timing of test positivity (Table 2.4), from the earliest detection of viraemia (≥100 copies/ml) (Fiebig et al. 2003; Fiebig et al. 2005). Fiebig staging is shown in Table 2.4 (Fiebig et al. 2003). The staging method does not estimate the date of possible transmission accurately within days of the true date (Fiebig et al. 2003; Fiebig et al. 2005). Considerable variability in the onset of the exponential phase of viraemia has been demonstrated (Fiebig et al. 2005). Pre-exponential, low-level viraemia detected up to one month prior to the exponential phase (Fiebig et al. 2005). This 'Pre-Fiebig' stage has been added to Tables 2.3 and 2.4, in addition to the original staging descriptions and estimates (Fiebig et al. 2003).

Table 2.4 Estimated timing and duration of Fiebig stages from the onset of detectable viraemia (day 0). *(Fiebig et al. 2003; Fiebig et al. 2005)*

Fiebig Stage	Cumulative duration in days (95% CI) from onset of detectable viraemia in the exponential growth phase
Pre-Fiebig	unknown, at least -31 days
I	0 to +5.0 (+3.1 to +8.1)
II	+5.1 to +10.3 (+7.1 to +13.5)
III	+10.4 to +13.5 (+10.0 to +17.0)

IV	+13.6 to +19.1 (+15.3 to +22.9)
V	+19.2 to +88.6 (+47.4 to +129.8)
VI	Chronic phase of infection with open-ended duration

2.2.3.2 Problems with using Fiebig staging to infer the estimated time of transmission

Some authors (Abrahams et al. 2009; Bar et al. 2010; Keele et al. 2008; Li et al. 2010; Salazar-Gonzalez et al. 2008) have inferred an approximate time of transmission by a using a method based upon Fiebig staging (Fiebig et al. 2003) and mathematical modelling of exponential growth based upon another rare cohort (Stafford et al. 2000). However, there are some important caveats to inferring the estimated time of transmission by this method. Firstly, as described, the Pre-Fiebig stage is of unknown, and possibly highly variable, duration (Fiebig et al. 2005). Secondly the mathematical model for exponential growth was itself based upon a small cohort where the timing of infection was inferred for most participants (Stafford et al. 2000). Thirdly, demographic information was unknown for the anonymized donors (Fiebig et al. 2003; Fiebig et al. 2005). Intravenous drug use, or sexual contact with intravenous drug users, was known to be common among seropositive blood donors (Williams et al. 1988). There is insufficient evidence as to whether intravenous drug users demonstrate a different duration of the pre-Fiebig stage compared with men who have sex with men. However, intravenous drug users would be more likely to have multi-variant infections (Bar et al. 2010; Li et al. 2010). Therefore, other aspects of transmission and early infection may also be quantitatively different.

2.2.3.3 Minimum Pre-Fiebig estimation

I used a conservative assumption of at least a five-day minimum interval between sexual transmission of HIV-1 and detectable viraemia by conventional methods (Chapter 3). I also assumed a minimum *in vivo* a generation time of between one and 2.5 days (Chapter 3), but I noted that these were inferred estimates from models (McLean and Nowak 1992; Perelson et al. 1996; Stafford et al. 2000). Therefore, unlike previous authors (Abrahams et al. 2009; Bar et al. 2010; Keele et al. 2008; Li et al. 2010; Salazar-Gonzalez et al. 2008), I have not used Fiebig staging and modelling-based assumptions for estimating maximum Pre-Fiebig interval between transmission and detectable viraemia.

2.2.3.4 Clinical assay sensitivity and specificity

Fiebig staging is specific to the sensitivity and specificity of the assays used in the original paper (Fiebig et al. 2003), which had been superseded by more sensitive assays in the cohorts presented in Chapters 3 to 6. Firstly, the minimum cut-off for detectable viraemia was higher (100 copies/ml) than the minimum detectable level for the cohorts that I analysed (50 copies/ml). Secondly, the protocols for the cohorts that I analysed made use of sensitive fourth generation ELISAs, which detect p24 Ag as well as Ab, to identify seroconverters (Pandori et al. 2009; Weber et al. 1998). Fiebig and colleagues did use a separate, dedicated p24 Ag assay for retrospective staging of the older of the two cohorts (Fiebig et al. 2003), but there are limited data to support the reliability of testing samples frozen for more than a decade at -20°C. Dedicated p24 Ag assays may have been more sensitive than some fourth-generation ELISA assays (Sickinger et al. 2004; Weber et al. 1998) but p24 antigenaemia is usually transient and may have been missed by

Fiebig and colleagues where there were gaps in samples (Fiebig et al. 2003). Furthermore, Fiebig and colleagues included data from seroconverters who were originally identified among blood donors using older assays (Fiebig et al. 2003; Williams et al. 1988), which either only detected antibody or were less sensitive than more modern screening assays (Stramer et al. 2004; Weber et al. 1998). Lastly, p31 Ag data were not available for all participants that I analysed, as either this test was not routinely included in WB kits used in Australia or WB was not routinely performed for UK seroconverters.

2.2.4 Ethics Statement

The SPARTAC Trial scientific studies were approved by the Multicentre Research Ethics Committee (MREC) and the use of materials from the Royal Berkshire Hospital, Reading, was approved by the Oxford Research Ethics Committee as part of the Thames Valley Cohort. All participants provided written informed consent before participating in these studies.

2.3 Laboratory Methods

2.3.1 Definitions and abbreviations

Table 2.5 Abbreviations

Term	Definition
1 bp	1 base pair
1 kb	1 kilobase
PCR	Polymerase chain reaction
SGA	Single genome amplification

2.3.2 Sampling of sequences within hosts

2.3.2.1 Single genome amplification

Single genome amplification (SGA) is a method of sampling sequences by diluting DNA template solution and then aliquoting the template solution across multiple wells so that a positive well probably only contained a single DNA template copy (Jordan et al. 2010; Palmer et al. 2005; Salazar-Gonzalez et al. 2008; Simmonds et al. 1990). A Poisson distribution is used to calculate this probability using the number of negative wells, because negative wells are assumed not to contain a cDNA template copy (Simmonds et al. 1990). The usual threshold for assuming single-copy amplification is 70% or higher negative wells (Jordan et al. 2010; Salazar-Gonzalez et al. 2008). The advantage of single genome amplification is that it reduces the risk of in vitro recombination due to PCR (Jordan et al. 2010; Meyerhans et al. 1990; Palmer et al. 2005). The disadvantages of SGA are that obtaining the correct dilution can be costly in terms of template and published studies generally do not include validation of PCR sensitivity (Jordan et al. 2010). In addition, SGA is no better than other sampling methods in terms of limiting either primer bias or template re-sampling (Jordan et al. 2010).

2.3.2.2 Bacterial cloning of population amplified template

Bacterial cloning is an efficient way of quantifying diversity after population-amplification PCR (Jordan et al. 2010). The chief drawback was that the risk of *in vitro* recombination at the population PCR step would invalidate the sampled viral sequence haplotype, especially in long fragments (Jordan et al. 2010; Salazar-Gonzalez et al. 2008).

2.3.2.3 In vitro *recombination with population-amplification PCR*

Previous authors have criticized the use of taq-based population-amplification sequencing because of the risk of taq-induced recombination *in vitro* (Salazar-Gonzalez et al. 2008), so the risk of PCR-induced recombination my have been exaggerated (Erlich et al. 1991). Many reported problems with PCR-induced recombination could be at least partly attributed to methodological errors in the PCR. Firstly, using a short PCR elongation step for a long target fragment will increase the rate of recombination. Some authors (Salazar-Gonzalez et al. 2008) have used a protocol with a PCR elongation step several minutes shorter than the manufacturer's recommended elongation time for the target fragment size (Roche 2005). Furthermore, enzymes are limited to reliably amplifying target fragments of a particular size. The target fragment amplified by previous authors (Salazar-Gonzalez et al. 2008) have been at the upper size limit for the enzyme used (Roche 2005). Lastly, increasing the number of PCR cycles will increase the risk of recombination (Meyerhans et al. 1990), but previous authors (Salazar-Gonzalez et al. 2008) have run the maximum permissible number of cycles for the enzyme used (Roche 2005) for two cycles of nested PCR, regardless of the initial template concentration, when

fewer cycles may suffice with sensitive primers (Jordan et al. 2010).

2.3.3 Steps taken to prevent contamination

Care was taken to keep genetic material from different plasma samples separate. In particular, samples from the same participant, or an epidemiologically linked participant, were handled separately to prevent contamination. Negative controls were run for each stage of sequence amplification. All reagents were aliquoted before use with genetic template. In addition, all sequences were checked on phylogenetic trees to ensure that epidemiologically unlinked sequences did not cluster together.

2.4 Analytical Methods

2.4.1 Abbreviations

Table 2.6 Abbreviations

Abbreviation	Term
1sKH test	One-sided Kishino-Hasegawa test
α	Alpha shape parameter
ELW	Expected likelihood weight
ESS	Effective sample size
Γ distribution	Gamma distribution
HD	Hamming Distance
I	Proportion of invariable sites parameter
LR	Likelihood Ratio
ML	Maximum likelihood
MCMC	Markov chain Monte Carlo
NNI	Nearest neighbour interchange
SGA	Single genome amplification
SH test	Shimodaira-Hasegawa test
ULRM	Uncorrelated lognormal relaxed molecular clock
UERM	Uncorrelated exponential relaxed molecular clock

2.4.2 Sequence alignment

2.4.2.1 Manual alignment

In this thesis, all nucleotide sequences were initially aligned manually in Se-Al v2.0a11 against the reference sequence HXB2 (Shaw et al. 1984). Codon structure was maintained for all coding regions.

2.4.2.2 Algorithm-based alignment

For multiple-sequence alignments of *env*, nucleotide sequences were aligned using MUSCLE v3.7 (Edgar 2004a, 2004b). Alignments generated by MUSCLE v3.7 (Edgar 2004a, 2004b) were manually edited in Se-Al v2.0a11 to restore codon structure in coding regions.

2.4.2.3 Pairwise analysis

All clonal and sequential nucleotide sequences were sorted by pairwise distance and visualized in the program Highlighter v2.1.1 (LANL, US). (Laboratory) All unique nucleotide sequences of supposedly viral origin were entered into the Basic local alignment search tool (BLAST), which compares sequences with the GenBank database and finds regions where sequences are similar. The BLAST search was used to check the raw similarity of all amplified sequences to database sequences of HIV-1.

2.4.2.5 Gap-stripping

Multiple nucleotide alignments with regions of poor alignment were gap-tripped using Gapstreeze (LANL, US) set to 5%. In GapStreeze, the user sets a gap tolerance between 0% and 100%. A value of 5% will cause all columns in the alignment to be deleted if more than 5% of sequences contain a gap at that position. Gap-stripped sequences were manually checked in Se-Al v2.0a11 to ensure that codon structure had been maintained in the alignment. Gap stripping ensured that the genetic distance was calculated for homologous regions only, so that the estimated distance was not inflated during phylogenetic tree reconstruction (Lemey et al. 2009).

2.4.2.6 Hamming distance

The Hamming distance (HD) was used to measure the degree of difference between two aligned, gap-stripped sequences. Clonal sequences or samples from the same participant taken at different time points. I calculated the HD by summing the number of positions in the aligned sequences that differed, and then dividing the total by the length of the alignment. This HD is often an underestimate of the true genetic distance because there is not

correction for multiple substitutions at a single locus or back mutations (Lemey et al. 2009). Therefore, I only used the HD as part of a raw screening test for recombination.

2.4.3 Phylogenetic analysis

2.4.3.1 Phylogenetic signal

Multiple nucleotide sequence alignments containing only clonal, SGA or sequential samples that were epidemiologically linked were checked for adequate phylogenetic signal. Also checked were sequences that were less than 250bp in length. I determined that there was adequate phylogenetic signal by analysing the number of phylogenetically informative sites. Phylogenetic signal was assessed by the quartet-puzzling method, as implemented in TREE-PUZZLE v5.3.rc7! (Schmidt et al. 2002; Schmidt and von Haeseler 2007).

2.4.3.2 Markov models

It is not currently possible to observe the actual process of mutation by point mutation in most real datasets. To implement phylogenetic analyses, this thesis uses the assumption of Kimura that independent point substitutions of nucleotides in nature occur stochastically in finite populations and that observed substitutions are selectively neutral (Kimura and Crow 1964; Kimura 1968). Therefore, mutation by substitution is modelled statistically as a Markov process (Lemey et al. 2009). Under a Markov model, the probability of any future state is conditioned only upon the present state, not upon previous states (Lemey et al. 2009). However, HIV-1 possesses a small genome that is almost entirely composed of coding regions or transcription factor binding regions, the evolution of which may violate the assumption of neutrality (Holmes 2009; Lemey et al. 2009). Therefore, the assumption of selectively neutral

substitutions has been used with caution, even where corrections were made for rate heterogeneity among sites and the proportion of invariable sites.

2.4.3.3 Further assumptions when applying Markov models

To apply a Markov model to observed data, further simplifying assumptions are necessary. Assuming that a Markov model of substitution estimates the true substitution process allows a researcher to infer genetic distances within a set of observed sequences (Lemey et al. 2009). The observed pairwise distances in these sequences are compared with expected genetic distance under the assumption of a particular Markov model (Lemey et al. 2009). It is assumed that the rate of substitution remains fixed over time and that the relative frequencies of nucleotides in the alignment are in equilibrium (Lemey et al. 2009). The latter can be checked statistically by comparing the observed nucleotide frequency in each sequence of a dataset with that expected under a particular Markov model (Lemey et al. 2009; Schmidt et al. 2002; Schmidt and von Haeseler 2007). Finally, it was also assumed that substitution is the only mechanism of mutation, since other mechanisms such as recombination, insertions or deletions are not included in the model (Lemey et al. 2009). However, since HIV-1 mutates frequently by all these mechanisms, this last assumption is also used with caution (Holmes 2009; Lemey et al. 2009).

2.4.3.4 Nested Markov models of substitution

The Markov models used in this thesis are the general time-reversible (GTR) model (Lanave et al. 1984a; Lanave et al. 1984b; Tavaré 1986) and other time-reversible models nested within the GTR model (Lemey et al. 2009). The more complexity that is

included within a model, the more the model includes mechanistic data regarding the actual process of evolution by nucleotide substitution. However, the ability to confidently estimate model parameters *a posteriori* from the observed data is limited by the amount of genetic information contained within the data(Lemey et al. 2009), hence the use of simpler, nested models in some places in this thesis.

2.4.3.5 The general time reversible (GTR) model

The GTR model (Lanave et al. 1984a; Lanave et al. 1984b; Tavaré 1986) is the most general of the time-reversible models of nucleotide substitution (Lemey et al. 2009). The GTR model Q matrix, using the notation of Lemey and colleagues (Lemey et al. 2009), is:

$$Q = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(b\pi_A + e\pi_C + f\pi_G) \end{bmatrix}$$

.

The GTR model uses eight free parameters: three unequal base frequencies (π_A, π_C, π_G), three transversion frequency parameters (*a*, *c*, *d*) and two transition frequency (*b*, *e*) parameters (Lemey et al. 2009). Although there are two additional parameters (π_T, f) in the Q matrix, these parameters are not included among the free parameters, since $\pi_A + \pi_C + \pi_G + \pi_T = 1$ and $a + b + c + d + e + f = 1$ (Lemey et al. 2009).

2.4.3.6 The Tamura and Nei 1993 (TN93) model

The TN93 model (Tamura and Nei 1993) is the next most complex nested model after the GTR model, with five free parameters (Lemey et al. 2009). The TN93 model allows unequal base frequencies as

well as different rates between transition types only (Tamura and Nei 1993). There is a single rate for transversion. This model has the parameters for the rates of different types of transition, written as κ_1 and κ_2, using the parameter names of BEAST v1.6.1 (Drummond and Rambaut 2007). In this thesis, parameters κ_1 and κ_2 were estimated from the data. The Q matrix for the TN93 model, slightly modifying the notation of Lemey and colleagues (Lemey et al. 2009), is:

$$Q = \begin{bmatrix} -\mu(\pi_C + \kappa_1\mu\pi_G + \mu\pi_T) & \mu\pi_C & \kappa_1\mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \mu\pi_G + \kappa_2\mu\pi_T) & \mu\pi_G & \kappa_2\mu\pi_T \\ \kappa_1\mu\pi_A & \mu\pi_C & -\mu(\kappa_1\pi_A + \mu\pi_C + \mu\pi_T) & \mu\pi_T \\ \mu\pi_A & \kappa_2\mu\pi_C & \mu\pi_G & -\mu(\pi_A + \kappa_2\mu\pi_C + \mu\pi_G) \end{bmatrix}$$

2.4.3.7 The Hasegawa, Kishino and Yano 1985 (HKY85) model

The HKY85 model (Hasegawa et al. 1985) is the next more simple used. HKY85 is nested within TN93, with five free parameters (Lemey et al. 2009). The HKY85 model allows unequal base frequencies as well as unequal substitution rates between transitions and transversions (Hasegawa et al. 1985). There is only a single transition versus transversion parameter, κ (Lemey et al. 2009). If $\kappa > 1$, then transitions are more common than transversions (Lemey et al. 2009). In this thesis, the parameter κ was estimated form the data. The Q matrix for the HKY85 model (Hasegawa et al. 1985) slightly modifying the notation of Lemey and colleagues (Lemey et al. 2009) to that of Yang and colleagues (Z. Yang et al. 1995), is:

$$Q = \begin{bmatrix} -\mu(\pi_C + \kappa\mu\pi_G + \mu\pi_T) & \mu\pi_C & \kappa\mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \mu\pi_G + \kappa\mu\pi_T) & \mu\pi_G & \kappa\mu\pi_T \\ \kappa\mu\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \mu\pi_C + \mu\pi_T) & \mu\pi_T \\ \mu\pi_A & \kappa\mu\pi_C & \mu\pi_G & -\mu(\pi_A + \kappa\mu\pi_C + \mu\pi_G) \end{bmatrix}$$

2.4.3.8 Kimura 2-parameter model

The Kimura 2-parameter model is a simple nested model that only includes transition and transversion parameters (Kimura 1980; Lemey et al. 2009). This model was only used for bootstrap analysis when screening for recombination breakpoints in conjunction with other tests (Lemey et al. 2009).

2.4.3.9 Rate heterogeneity among sites

Where rate heterogeneity as considered in this thesis, α was estimated from the data. The assumption of neutrality is relaxed partially by the addition of a parameter for rate heterogeneity among sites, particularly in gene coding regions (Lemey et al. 2009). Rate heterogeneity among sites is estimated to occur according to a Γ distribution, with a shape parameter, α, that can be set *a priori* or estimated *a posteriori* (Lemey et al. 2009). A substitution model nominally 'without' an additional α parameter in fact a special case of $\alpha \to \infty$ (Z. Yang et al. 1995). The special case $\alpha \to \infty$ was also used as an analytical control for analyses involving only epidemiologically linked infections. This was because apparent rate heterogeneity among sites may be an artefact of the high degree of relatedness of between the sequences, rather than reflecting true rate heterogeneity.

2.4.3.10 Proportion of invariable sites

I included an invariable sites parameter estimate for phylogenetic analyses involving sequences from epidemiologically unlinked infections. This parameter partially accounts for site invariance due

to non-neutral mutation in the model (Lemey et al. 2009; Steel et al. 2000). In HIV-1, this may be due to differences in epitope targeting by the immune system of different individuals, for example. The proportion of invariable sites, I, in an alignment can be set *a priori* or estimated *a posteriori* and $0 \leq I \leq 1.0$ (Lemey et al. 2009; Steel et al. 2000). A substitution model nominally 'without' an additional I parameter is in fact a special case of $I = 0$ (Lemey et al. 2009; Steel et al. 2000). I did not apply the invariant sites parameter to alignments containing only sequences from epidemiologically linked infections, as the proportion of invariant sights could be high due to the recent common ancestry of the sequences, rather than because the sites were evolutionarily invariant due to changes in the non-neutrality of substitutions (Lemey et al. 2009; Steel et al. 2000).

2.4.3.11 Model selection

Selecting the optimal model for a given dataset is important because accurate estimation of one parameter affects the accurate estimation of other parameters in the model (Lemey et al. 2009; Z. Yang et al. 1995). This is true both for maximum likelihood estimation as well as Bayesian MCMC based estimation (Lemey et al. 2009). For example, estimation of the differences in base frequencies in the dataset is important for estimating the shape parameter α for rate heterogeneity among sites and the transition/transversion ratio κ (Z. Yang et al. 1995). Therefore, the choice of model is important for accurate estimation of branch lengths and, sometimes, the topology of the tree (Lemey et al. 2009; Z. Yang et al. 1995). However, increasing parameterization by using a more complex model will also increase the degree of variance in parameter estimates.

2.4.3.12 Methods of model selection

In this thesis models have been selected using either jModelTest (Posada 2008, 2009) or PAUP*v4.0 (Sinauer, US. The aim was to use the most biologically realistic model for which parameters could be estimated with confidence from the observed data (Lemey et al. 2009; Z. Yang et al. 1995). If there was a strong alternative explanation for the proportion of invariant sites, other than non-neutrality of substitutions, this parameter was not included regardless of the output from these two programs. Furthermore, if there was uncertainty regarding how many biological parameters could be estimated from an alignment containing only highly similar sequences, a number of nested models were tested to see if there was any effect on branch length estimates.

2.4.3.13 Tree topology

Tree topology was set *a priori* or trees were sampled using at least two different methods (Z. Yang et al. 1995), with the final tree compared with other information derived independently of the sequence dataset (Lemey et al. 2009). The tree topology is crucial because estimations of branch lengths rely on an accurate tree topology. Tree topology is not a parameter in a phylogenetic substitution model (Z. Yang et al. 1995). Instead, the parameters of the substitution model are estimated from a particular topology.

2.4.3.14 Co*nfidence in the estimated tree topology*

Confidence in the analytically estimated tree topology is challenged by the computational difficulty of sampling tree topologies that are possible for a large dataset with n samples (Lemey et al. 2009). The number of possible rooted, bifurcating tree topologies for $n \geq 2$ is:

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

and the number of possible unrooted, bifurcating tree topologies for $n \geq 3$ is:

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}.$$

This thesis used a Bayesian MCMC approach to tree sampling in addition to the nearest-neighbour interchange with non-parametric bootstrapping method implemented in the maximum-likelihood approaches (Guindon and Gascuel 2003; Huelsenbeck and Ronquist 2001; Lemey et al. 2009; Ronquist and Huelsenbeck 2003; Ronquist et al. 2012). For large alignments ($n > 50$) the faster maximum likelihood approach was used before the computationally expensive Bayesian MCMC approach.

2.4.3.15 Using outlier sequences to root a tree

Where the tree topology was derived analytically, rather than set *a priori*, multiple reference sequences were used in this thesis as outliers to root the tree. For example, if two UK subtype B sequences were being compared, it was assumed that other Group M or non-M group subtypes have branched earlier (Lemey et al. 2009). Furthermore, it was assumed that if epidemiologically unlinked infections separated two sequences within the UK epidemic, then the two sequences being analysed were also unlinked (Pillay et al. 2007). However, if two sequences clustered together to the exclusion of all other sequences, this did not 'prove' that the sequences were epidemiologically linked (Hayman et al. 2001; Pillay et al. 2007). In this thesis, reference sequences were drawn from the same cohort as the sequences being considered, as well as from the Los Alamos

National Laboratory (LANL) HIV sequence database.

2.4.3.16 Maximum likelihood estimation

I used maximum likelihood (ML) based estimation to make initial phylogenetic inferences. The ML tree is the tree that maximizes the likelihood of the observed data, given the tree topology and the substitution model (Lemey et al. 2009). Likelihood ratios were computed in PhyMLv3.0 (Guindon and Gascuel 2003). PhyML v3.0 (Guindon and Gascuel 2003) starts with a complete distance-based neighbour-joining (NJ) tree and then samples trees using a nearest neighbour interchange (NNI) algorithm. Parameters are estimated for each topology until a ML estimate is reached (Guindon and Gascuel 2003). Then the topology undergoes NNI and further likelihood maximization. The analysis terminates when no additional NNI can further maximize the ML. PhyML v3.0 is faster than Bayesian MCMC based approaches, such as that implemented in Mr Bayes v3.1.2 or v3.2, with substantially smaller demands on finite computer clock speed. PhyML v3.0 is therefore very useful for initial phylogenetic inferences. However, tree-sampling algorithms may become stuck at a local ML, rather than finding the global ML (Lemey et al. 2009).

2.4.3.17 Bayesian MCMC based methods

Finding the global ML ensures the estimated tree is as close as possible to the true tree (Lemey et al. 2009), so this thesis also uses Bayesian MCMC based methods for phylogenetic inferences involving large datasets. A Bayesian MCMC based approach to tree sampling produces a probability distribution of trees (Lemey et al. 2009), from which a consensus tree may be constructed that represents the average over a sampled set of plausible trees (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The Bayesian MCMC based approach stands in contrast to ML approaches, which produce a single tree with a maximized likelihood compared to the other sampled trees. In this thesis, the program MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) or MrBayes v3.2 (Ronquist et al. 2012) was used for phylogenetic analyses in this thesis where a molecular clock was not enforced. Analytic convergence was reached when all estimated sample size (ESS) estimates were >300, because a high ESS reflects a low degree of correlation among samples (Drummond et al. 2002; Lemey et al. 2009). The major disadvantage for using a Bayesian MCMC based approach to phylogenetic inference is that trees may take several weeks to achieve convergence in MrBayes v3.1.2 or v3.2, while ML estimation for the same dataset may take minutes in PhyML v3.0.

2.4.4 Time of most recent common ancestor (tMRCA) estimation

2.4.4.1 Molecular clock

Estimation of the tMRCA requires the assumption of clock-like behaviour in a sequence alignment. Phylogenetic trees with free substitution rates, such as those obtained with PhyML v3.0 and MrBayes v3.1.2 or v3.2, did not assume clock-like behaviour (Lemey et al. 2009). Therefore, a different approach was needed for tMRCA estimation in this thesis.

2.4.4.2 Molecular clock rate

The rate of the molecular clock is determined by the nucleotide substitution rate (Lemey et al. 2009). In relatively conserved sequences, the substitution rate is relatively slow. In relatively less conserved sequences, the substitution rate is relatively quick. Furthermore, the rate of nucleotide substitution in the same genomic region may vary over time, especially in coding regions (Drummond et al. 2006; Lemey et al. 2009). This variation may be caused by changes in the type and strength of natural selection applied to that genomic region or replication rate (Drummond et al. 2006; Lemey et al. 2009). In HIV-1, for example, viral genes come under varying degrees of selection from the host immune system (Lemey et al. 2006) and the replication rate may be affected by the stage disease (Edo-Matas et al. 2011). Furthermore, changes in the size of a population, such as the profound genetic bottleneck encountered during HIV-1 transmission (Abrahams et al. 2009; Bar et al. 2010; Edwards et al. 2006; Frater et al. 2006; Kearney et al. 2009; Keele et al. 2008; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009; L. Q. Zhang et al. 1993; Zhu et al. 1993), can impact upon the observed evolutionary rate (Drummond et al. 2006;

Edwards et al. 2006; Lemey et al. 2006; Lemey et al. 2009).

2.4.4.3 Strict molecular clock

A strict molecular clock assumes that the substitution rate is constant among all branches in a tree (Drummond et al. 2006; Kimura 1968, 1969; Lemey et al. 2009; Zuckerkandl and Pauling 1962, 1965a, 1965b). The strict molecular clock fits data well if the evolutionary rate remains constant over time (Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009). The strict molecular clock is a special case of the relaxed molecular clock, in which the substitution rate variation parameter, also called the evolutionary rate parameter, is set to 1.0 (Drummond et al. 2006; Lemey et al. 2009).

2.4.4.4 Relaxed molecular clock

A relaxed molecular clock retains the assumption of clock-like behaviour but allows for more realistic variations in evolutionary rate among lineages, rather than assuming that all lineages must evolve at an equal and constant rate (Lemey et al. 2009). Variation in the substitution rate may occur if evolution occurs at a different rate among the branches of a tree due to changes in selection (Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009). However, a relaxed molecular clock model remains useful for quantifying genetic divergence in terms of the tMRCA (Drummond et al. 2006; Lemey et al. 2009). Although a relaxed clock is a more realistic model of evolution over time, applying it to real datasets leads to an increase in the variance of the tMRCA estimate, and thus widens the confidence intervals for the estimated tMRCA (Edo-Matas et al. 2011; Lemey et al. 2006; Lemey et al. 2009). In particular, the uncorrelated lognormal molecular clock will have a long tail with

respect to the lower 95% CL for the tMRCA estimate because of a skew towards a faster evolutionary rate among more divergent sequences.

2.4.4.5 Uncorrelated lognormal relaxed molecular (ULRM) clock

A ULRM clock was used for analyses involving a relaxed molecular clock. Under a ULRM clock, the rate variation among branches is estimated by assuming a lognormal distribution (Drummond et al. 2006; Lemey et al. 2009). In addition, the model is uncorrelated because the rate of a branch is not dependent upon the rate of adjacent branches (Drummond et al. 2006; Lemey et al. 2009). Other models of a relaxed molecular clock are an uncorrelated exponential relaxed molecular (UERM) clock and a random local clock. A UERM clock is similar to a ULRM clock except that rate variation among branches is estimated by assuming an exponential distribution (Drummond et al. 2006; Lemey et al. 2009).The ULRM clock was used because it has been shown to give more useful estimates, with narrower confidence intervals, than the UERM clock. The UERM clock provides less useful estimates than the ULRM because it does not approximate biological clock-like behaviour as well as the ULRM clock does (Drummond et al. 2006).

2.4.4.6 Dynamic population size and coalescence model

It has been inferred that HIV-1 undergoes a profound genetic bottleneck at transmission (Abrahams et al. 2009; Bar et al. 2010; Edwards et al. 2006; Frater et al. 2006; Kearney et al. 2009; Keele et al. 2008; Li et al. 2010; Salazar-Gonzalez et al. 2008; L. Q. Zhang et al. 1993; Zhu et al. 1993). This bottleneck may occur because the successfully transmitted virus following sexual exposure represents only a tiny minority of all circulating viruses within a donor (Bar et

al. 2010; Keele et al. 2009a; Li et al. 2010; Salazar-Gonzalez et al. 2008). Therefore, the assumption of constant effective population size was relaxed in this thesis for coalescence modelling.

2.4.4.7 Bayesian skyline plot

The constant population size assumption was relaxed by implementing the Bayesian skyline method for all analyses (Drummond et al. 2005). For example, one would expect a period of viral population expansion within the host following transmission, and may be included in the model when modelling changes in early within-host viral diversity. The Bayesian skyline plot tool frees the user from having to enforce assumptions regarding population size dynamics, such as an exponential growth curve, when analysing real-world data for which these parameters are poorly characterized (Edwards et al. 2006). An alternative approach to the Bayesian Skyline plot is to fit the population change assumption to a particular curve. For example, an exponential growth curve is assumed in the Poisson-Fitter tool, which is explicitly a simplification of Bayesian methods (Keele et al. 2008; H. Y. Lee et al. 2009). The advantage of the Poisson Fitter is that it does not require parameter values for population size dynamics to be derived from the data, as this may not currently be possible with sets of near-identical sequences.

2.4.4.8 Bayesian MCMC based approach

Available prior information was specified before the analysis, using BEAUti v1.6.1 (Drummond and Rambaut 2007). Plausible trees were then sampled using a Bayesian MCMC based approach to simultaneously estimate posterior distributions for all parameters in the model. Convergence was reached when all ESS estimates were

>300 because a high ESS reflects a low degree of correlation among samples (Drummond et al. 2002; Lemey et al. 2009). A consensus tree was constructed from the plausible trees that had been sampled.

2.4.5 Recombination analysis

2.4.5.1 Overview

Recombination analysis was carried out to detect sequences from epidemiologically unlinked, same-subtype (intra-subtype) sequences circulating within the same host. Due to the high degree of similarity within subtypes, a number of methods were used to ensure that the result was robust to the choice of detection method (Casado et al. 2007; Gottlieb et al. 2004; Pacold et al. 2010; van der Kuyl and Cornelissen 2007).

2.4.5.3 Screening for recombination by query-versus-reference

Firstly unstripped reference sequences were identified based on genetic distance in the ML analysis. All query unstripped sequences were then visualized in Highlighter v2.1.1 (LANL, US). All sequences were screened using the sliding window-based methods implemented in RIP v3.0 (LANL, US), and the results were confirmed using SimPlot v3.5.1 (Lole et al. 1999).

2.4.5.3 Phylogenetic sliding window methods

Bootstrap support for the result obtained by the sliding window-based result obtained in SimPlot v3.5.1 (Lole et al. 1999) was confirmed by bootstrapping (1000 replicates). A simple tree-building (neighbour-joining) and substitution model (Kimura 2-parameter) algorithm was implemented in SimPlot v3.5.1 (Lemey et al. 2009).

2.4.5.4 Mosaic triplet methods

Confirmation of approximate breakpoint positions obtained by the

previous methods was confirmed using a mosaic triplet-based algorithm, as implemented in 3Seq (Boni et al. 2007; Hogan and Siegmund 1986; Lemey et al. 2009). The mosaic triplet-based algorithm does not depend upon the user identifying the query and reference sequences beforehand (Boni et al. 2007; Hogan and Siegmund 1986; Lemey et al. 2009). All sequences in the alignment were re-analysed by this method.

Chapter 3 Inferring the contemporaneous transmission of divergent viral variants from a single donor to two recipients

Parts of this chapter were published in: English, S. Katzourakis, A. Bonsall, D. Flanagan, P. Duda, A. Fidler, S. Weber, J. McClure, M. SPARTAC Trial Investigators, Phillips, R. Frater, J. (2011), 'Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants', *Retrovirology,* 8, 54.

3.1 Overview

3.1.1 Brief summary of chapter and hypothesis

The finding in this chapter was that the genetic divergence of HIV-1 strains that were contemporaneously transmitted from a single donor host could not be explained by immune reactions within the recipient hosts. I initially hypothesized that recipient human hosts who had been contemporaneously infected by a single donor would have very similar sequences. To test this hypothesis, I studied a rare case study of probable contemporaneous transmission of HIV-1 from a single, unsampled donor to two sampled recipients on the same evening. Firstly, I evaluated the clinical data for evidence of the timing of infection to ensure consistency with the reported single exposure to a common donor. Secondly, I used phylogenetic methods to demonstrate that the sampled recipients were infected with highly related viral variants. Despite being highly related, the *env* and *nef* variants had considerably more between-host divergence compared with the within-host divergence and I went on to quantify this divergence in Chapter 4 in terms of time to the most recent common ancestor. Thirdly, given the viral genetic divergence

between-hosts, I evaluated the humoral and cytotoxic immunological data from both participants for evidence of hard selection around the time of transmission. I also noted that there appeared to be more than one *pol* variant in either host, despite sequence homogeneity in the other genes studied, especially *env*. I concluded that there was no evidence of hard selection by the immune systems of either participant that would account for the genetic divergence between the viral variants infecting each host. The possible transmission of multiple *pol* variants also suggested that transmission of single variants was unlikely but there was no evidence to reject a random model of divergent variant transmission. Thus transmission of HIV-1 and the establishment of infection in a new host would appear to be a more complicated process than that assumed by the single variant transmission model.

3.1.2 Key abbreviations

Table 3.1 Abbreviations

Abbreviation	Term
bp	Base pair(s)
CCR5	C-C chemokine Receptor type 5
cDNA	Complementary DNA
CXCR4	C-X-C chemokine Receptor type 4
E. coli	*Escherichia coli*
EDTA	EthyleneDiamineTetraacetic Acid
ELISA	Enzyme-Linked ImmunoSorbent assay
ELISpot	Enzyme-Linked ImmunoSpot
ESS	Effective Sample Size
FCS	Fetal Calf Serum
x g	Times gravity
$+\Gamma$	Gamma distribution for rate heterogeneity among sites
GTR	General time reversible
LANL	Los Alamos National Laboratory
MCMC	Markov Chain Monte Carlo
pDNA	Plasmid DNA
PBMC	Peripheral Blood MononucleoCyte
PCR	Polymerase Chain Reaction
PNLGS	Potential N-linked glycosylation site
RT	Reverse Transcriptase
SGA	Single Genome Amplification
STN	Signal-to-Noise

3.2 Introduction

3.2.1 Antigenicity and vaccine design

A successful HIV-1 vaccine would be designed based upon the antigenicity of transmissible viruses. At the global level, multiple subtypes with evidence of on-going evolution (Kawashima et al. 2009) result in a level of diversity that has already frustrated all efforts to synthesize a universal HIV-1 vaccine (Gaschen et al. 2002). Additionally, substantial virus diversity develops within a single host during chronic infection (Shankarappa et al. 1999) and it is unclear which viral variants are transmissible to a new host. Recent efforts have concentrated on inferring variant transmissibility by characterizing the precise genetic and antigenic features of viruses found during very early stages of infection (Derdeyn and Hunter 2008; Edwards et al. 2006; Keele et al. 2008; Ochsenbauer et al. 2012; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009).

3.2.2 Genetic bottleneck

Single viral variants dominate in the majority of new, sexually-transmitted HIV-1 infections *in vivo*, indicating a profound genetic bottleneck around the time of transmission (Abrahams et al. 2009; Keele et al. 2008; H. Li et al. 2010). The degree of genetic bottleneck has been associated with the type of transmission mucosa (Abrahams et al. 2009; Bar et al. 2010; H. Li et al. 2010). Another factor associated with the number of infecting variants is the presence of mucosal infections (Haaland et al. 2009). Previous researchers have also suggested that differences in the degree of genetic bottleneck are related to variations in mucosal defence and its integrity (Haase 2005).

3.2.3 Possible confounding factors

However, the actual mechanism of this genetic bottleneck remains unclear and studies may be confounded by variations in both the risk of transmission among donors and the diversity of transmissible virions within donors (Edwards et al. 2006). The highest risk of transmission occurs during primary infection when the population size of infectious virus peaks (Hollingsworth et al. 2008). However, viral diversity within the acutely infected donor is limited, potentially making transmitted viruses indistinguishable in the recipient (Abrahams et al. 2009; Keele et al. 2008; Lemey et al. 2006; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009).

3.2.4 The genetic bottleneck may not be caused by the mucosal barrier

Furthermore, genetic analysis has also indicated that mucosal defence and integrity are not the only explanations for the apparent genetic bottleneck. Demographic models have been developed that avoid unsupported prior assumptions about the degree of genetic bottleneck (Drummond et al. 2005). Viral variability was compared (Edwards et al. 2006) in *gag* and *env* genes after transmission in mother-to-child transmission cases and in men who have sex with men (MSM). Viral variability after transmission was not consistently associated with the route of transmission (Edwards et al. 2006). In addition, a severe genetic bottleneck may be a sufficient, but not a necessary, condition for random transmission of genetic variability (Edwards et al. 2006).

3.2.5 Envelope proteins and transmissibility

Other authors have hypothesized that if transmission of viral

variability is not random, then traits which enhance transmission may be an adaptation (L. Q. Zhang et al. 1993; T. Zhu et al. 1993). However, transmissibility has not yet been associated with specific viral characteristics. Most new, sexually-transmitted HIV-1 infections are CCR5-tropic (Butler et al. ; Salazar-Gonzalez et al. 2009), but this may reflect biased representation of these variants in genital fluids (Bull et al. 2009; Diem et al. 2008). In eight cases of heterosexual transmission of subtype C (Derdeyn et al. 2004), transmitted variants tended to have fewer potential N-linked glycosylation sites (PNLGSs) and shorter hypervariable loops than the average variant in the donor host. In addition, recipient env-pseudotyped virus was more susceptible to neutralization by donor serum than donor env-pseudotyped virus (Derdeyn et al. 2004). A study of 35 subtype A cases from Kenya and 13 subtype B cases from the USA (Chohan et al. 2005) found that recently-infected persons had viruses with shorter, less-glycosylated V1V2 loops compared with a database of viruses (Chohan et al. 2005). However, studies of subtype B have not shown a consistent decrease in hypervariable loop length or the number of PNLGSs (Frost et al. 2005; Y. Liu et al. 2008). Therefore, there is no firm evidence that natural selection determines transmission of viral variants.

3.2.6 Comparisons with an animal model

Animal models of HIV infection that use the closely-related simian immunodeficiency virus (SIV) have also demonstrated that many different variants circulating within the host are transmissible. A low-dose, intrarectal inoculum of SIV was given to 18 rhesus macaques (Keele et al. 2009a) to mimic physiological concentrations. Although between one and five variants initiated new infections, the

viruses transmitted to all macaques collectively reflected the diversity within the inoculum (Keele et al. 2009a). Another study (Rybarczyk et al. 2004) demonstrated a stochastic pattern of *V1V2* variant transmission from an inoculum. Therefore a broad range of viruses circulating in a single donor may be potentially transmissible at any one time, consistent with the hypothesis that successful transmission of viral variants is a random process.

3.2.7 Aims

To investigate whether HIV-1 transmission is also stochastic in humans, I investigated a rare case study of contemporaneous HIV-1 transmission from a single donor to two recipients. The aim of my study was to see if hard selection by the immune system could account for the apparent transmission of divergent variants.

3.3 Materials and Methods

3.3.1 Participants

Two male individuals from the SPARTAC UK cohort (Chapter 2), P1 and P2, were identified on clinical history as having epidemiologically linked infections: they were partners and had shared a sexual encounter with a single, third male on the same night. P1 and P2 were enrolled in the SPARTAC trial on the same day and followed up at the Jefferiss Trust Clinic, St. Mary's Hospital, Paddington, London, UK. They were both randomized to receive no therapy.

3.3.2 Laboratory Methods

3.3.2.1 Separation of PBMCs and plasma

PBMC and plasma samples were separated from fresh EDTA blood according to previously published methods (Duda et al. 2009; Frater et al. 2007).

3.3.2.2 Viral RNA extraction

1ml aliquots of frozen plasma were used for each extraction. The plasma was centrifuged at 1600 x *g* and 4°C for 1 h to pellet the virus. Excess plasma was removed and the pellet was re-suspended in 140μl of remaining plasma. RNA was then extracted with the QIAamp Viral RNA Minikit (Qiagen, UK) according to the manufacturer's instructions.

3.3.2.3 Reverse transcription

For *env* and *nef*, viral RNA was reverse transcribed using the SuperScript III Kit (Invitrogen, UK) to produce cDNA. 15μl of viral RNA was added to 1.5μl dH_2O, 1.5μl primer OFM19 (Keele et al. 2008) (concentration 20μM) and 1.5μl dNTPs (concentration 10mM).

The mix was heated to 65˚C for 5 min followed by 4˚C for 1 min to anneal the primers to the RNA. The RT reaction mix (5x Buffer: 6μl, DTT: 1.5μl; RNaseOUT 1.5μl; SuperScript III 1.5μl) was then added to make a final volume of 29μl. The reaction mix was heated to 50˚C for 60 min, followed by 55˚C for 60 min and finally 75˚C for 10 min.

For *gag* and *pol*, viral RNA was reverse transcribed using the Reverse-iT 1st Strand Synthesis Kit (Abgene, UK). 18μl of viral RNA was added to 1.5μl primer (random decamers and oligodT supplied with the kit, concentration 20μM). The mix was heated to 75˚C for 5 min followed by 4˚C for 2 min to anneal the primers to the RNA. The RT reaction mix (5x Buffer: 6μl; dNTPs: 3μl concentration 10mM; RTase Blend 1.5μl) was then added to make a final volume of 30μl. The reaction mixture was heated to 42˚C for 60 min followed by 75˚C for 10 min.

3.3.2.4 Population-amplification

I amplified HIV-1 *env* genes by PCR using a protocol for SGA as described in detail elsewhere (Keele et al. 2008; Salazar-Gonzalez et al. 2008) and as per the manufacturer's instructions (Invitrogen UK). Briefly, 1μl cDNA was added to 19μl of reaction mix (14.1μl dH_2O, 2μl 10x Buffer, 0.8μl Mg_2SO_4, 0.4μl dNTPs, 0.8μl Primer 1 (concentration 20μM), 0.8μl Primer 2 (concentration 20μM), and 0.1μl High Fidelity *taq*). The reaction mixture was heated to 94˚C for 15s and then entered cycles of the reaction (94˚C for 15s, 50˚C for 30s and then 68 ˚C for 5 or 4 min, depending on whether it was generating the first or second round PCR product, respectively). 35 cycles were used for the first round PCR reaction and 45 cycles were used for the second round PCR reaction. Finally the reaction was heated to 68˚C for 10 minutes and then held at 4˚C until removal to

long-term -20˚C storage. The HIV *gag* and *pol* genes were amplified by separate PCR reactions (Duda et al. 2009; Frater et al. 2007). I re-amplified the *gag* fragment to sequence a larger region using the same protocol (Duda et al. 2009; Frater et al. 2007). Briefly, 2μl of cDNA was added to 38.3μl of dH_2O, 5μl 10x Buffer, 2μl $MgCl_2$, 1μl dNTPs, 1μl Primer 1 (concentration 10μM), 1μl Primer 2 (concentration 10μM), and 0.2μl Platinum *taq*). The reaction mixture was heated to 94˚C for 15s and then entered cycles of the reaction (94˚C for 15s, 55˚C for 30s and then 72 ˚C for 3 or 2.5 min, depending on whether it was generating the first or second round PCR product, respectively). 30 cycles were used for the first and second round PCR reaction. Finally the reaction was heated to 72˚C for 7 minutes and then held at 4˚C until removal to long-term -20˚C storage. I amplified the *nef* fragment by PCR using a protocol described in detail elsewhere (Duda et al. 2009; Frater et al. 2007) and using the same concentrations as well as amplification conditions as for the long *gag* fragment. For a discussion of the merits of these amplification conditions, please see 2.3.2.3.

3.3.2.5 Single genome amplification (SGA)

I carried out SGA of *env* as described by previous authors (Keele et al. 2008; Salazar-Gonzalez et al. 2008), with a 70% minimum cut-off for well negativity (Keele et al. 2008; Palmer et al. 2005; Salazar-Gonzalez et al. 2008). I diluted baseline cDNA in water, according to the baseline viral load, and amplified a test strip of sixteen wells. I aimed to have 12-13/16 negative wells and adjusted the dilution accordingly. I then proceeded to amplify a 96-well plate. Amplification products were run on a 1% agarose gel (Sigma, UK). Bands of the expected size were extracted under UV visualization

and then the amplified DNA was purified using the QIAquick Gel Extraction Kit (Qiagen, UK) according to the manufacturer's instructions.

3.3.2.6 Bacterial cloning

Bacterial cloning was carried out using the TOPO TA "One Shot" Cloning Kit for Sequencing according to the manufacturer's instructions (Invitrogen, UK). Bacteria were lysed and minipreps of clonal pDNA were prepared using either the Montage Miniprep$_{96}$ Kit (Millipore, US) or the QIAprep Spin Miniprep Kit (Qiagen, UK) according to the manufacturer's instructions.

3.3.2.7 Sanger Sequencing

Sequencing of population PCR, SGA and bacterial cloning DNA products was performed using BigDye technology in a 96-well plate. For population PCR and SGA products, 3μl DNA was added to a mix containing 0.8μl BigDye Terminator (Applied Biosystems, UK), 1.5μl 5x sequencing buffer (Applied Biosystems, UK), 2μl of primer (3.3μM) and 2.7μl dH$_2$O. For bacteria-cloned pDNA, 4μl of miniprep was added to a mix containing 1μl BigDye Terminator, 1.5μl 5x sequencing buffer, 1μl of primer (3.3μM) and 3.5μl dH$_2$O. The following cycling conditions were used: 96˚C for 30 s, then 30 cycles of 96˚C for 30 s, 50˚C for 15 s and 60˚C for 4 min. DNA for sequencing was precipitated on ice with 2μl 3M sodium acetate, 10μl dH$_2$O, 50μl ice-cold 100% ethanol for 5 min at -20˚C, centrifuged at 600 x *g* for 80 min at 4˚C, washed twice with ice-cold 70% ethanol and run on an ABI 3700 sequencer.

3.3.3 Analysis methods

3.3.3.1 Sequence alignment

All sequences were manually edited using Sequencher v4.8 (Gene

Codes Corporation, US) and manually aligned using Se-Al v2.0a11 (Rambaut 2002). For *env* alignment, sequences were first aligned with MUSCLE v3.7 (Edgar 2004a, 2004b) followed by manual alignment. Sequences containing stop codons or frameshifts were deleted prior to subsequent analysis. Where appropriate, reference sequences were obtained from the LANL HIV sequence database. For *env*, which contains many gaps and poorly aligned regions, gap stripping was undertaken first with GapStreeze set to 5%. Sequences were manually edited in Se-Al v2.0a11 before and after gap stripping.

3.3.3.2 Between-host phylogenetic analysis

Phylogenetic analysis of viral sequences sampled from P1 and P2 was carried out by several methods across the *env*, *gag*, *pol*, and *nef* genes or fragments. Prior to gap-stripping with GapStreeze, a likelihood mapping analysis (Strimmer and von Haeseler 1997) was run to ensure phylogenetic signal within *env* was significant. Likelihood mapping was implemented in TREE-PUZZLE v5.3.rc7! (Schmidt et al. 2002; Schmidt and von Haeseler 2007) and the *env* fragment was screened from the beginning of the *gp160* coding start region to the end of *gp120* (HXB2 nucleotide position 6225 to 7757). Additionally, full nucleotide sequences for the fragments from all three genes were visually screened in Highlighter (LANL, US) and the inferred protein sequence were screened visually using Jalview v2.6 (Clamp et al. 2004; Waterhouse et al. 2009).

3.3.3.3 Phylogenetic tree estimation using ML

Phylogenetic trees were initially constructed using the ML method as implemented in PhyML v3.0 software (Guindon and Gascuel 2003) and visualized in FigTree v1.3.1 (Rambaut 2009). I chose the

substitution model that gave the highest likelihood with PAUP*v4.0 (Sinauer Associates, US), which was the GTR+Γ+I model. ML branch support values were obtained by non-parametric bootstrapping using PhyML v3.0 (Guindon and Gascuel 2003) with 1000 replicates (Hedges 1992; Van de Peer et al. 2000; Zharkikh and Li 1992). An ML approach was used to estimate per cent divergence, as implemented in TREE-PUZZLE v5.3.rc7! (Schmidt et al. 2002; Schmidt and von Haeseler 2007).

3.3.3.4 A Bayesian MCMC based method for estimating trees

Finally, phylogenetic analysis was repeated using a Bayesian MCMC based method, as implemented in MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). An unconstrained branch length (exponential) prior was used to avoid enforcing a molecular clock (Lemey et al. 2009). MrBayes v3.1.2 or v3.2 was run in duplicate for at least 50,000,000 steps for *env*, *pol*, the *nef*-only fragment and the large *gag* fragment, sampling trees every 1,000 steps. MrBayes v3.1.2 was run in duplicate for at least 100,000,000 steps for the small *gag* fragment, sampling every 10,000 steps. MrBayes v3.2 was run in duplicate for at least 110 million steps for the long *nef* fragment, sampling every 1,000 steps. Convergence was assessed with Tracer v1.5 (Rambaut and Drummond 2009) with all parameter estimates having ESSs of >300 (Lemey et al. 2009). The consensus tree for each gene, with posterior probabilities for branch support, was generated and visualized in FigTree v1.3.1.

3.3.3.5 Poisson-Fitter

Poisson Fitter is an algorithm based upon inferred values of virus population expansion and mutation rate during acute and very early infection (E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009; Stafford et al. 2000). The primary purpose of this tool is to estimate the number of successfully transmitted variants from the donor to the recipient (Abrahams et al. 2009; E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009), assuming that most homogeneous infections involve transmission of a single variant followed by neutral evolution in the absence of an effective immune response. The model is a simplification of coalescent methods (Drummond et al. 2002) because sequences with very limited

diversity do not contain sufficient signal for parameter estimation from the sequences alone (Keele et al. 2008; H. Y. Lee et al. 2009). Therefore, a number of parameters are fixed based on values obtained from other datasets (Keele et al. 2008), such as the nucleotide substitution rate (Mansky and Temin 1995) and the rate of viral population growth per day (Markowitz et al. 2003; Stafford et al. 2000). The user must also enter the sequence consensus, which I constructed in Se-Al v2.0a11. The consensus is assumed to approximate the founder virus sequence and the recombination rate is set to zero (Abrahams et al. 2009; E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009; Ochsenbauer et al. 2012). The change in Hamming Distance over time is then fitted to a Poisson distribution that incorporates these fixed parameter values, with this distribution constituting the null-hypothesis. Poisson-Fitter then uses an arbitrary cut-off of (for example, >10%) deviation between the expected and observed Hamming distance values to determine whether or not the sampled sequences conform to a star-like phylogeny.

3.3.4 Immunological profiling

3.3.4.1 Potential N-linked glycosylation site analysis

We compared potential N-linked glycosylation sites (PNLGSs) between inferred amino acid sequences for the SGA samples *env* in from P1 and P2, using N-Glycosite (LANL, US).

3.3.4.2 Neutralization and infectivity assays

HIV *env* genes were amplified from reverse transcribed viral RNA, restriction-cloned in pcDNA3.1 (Invitrogen, UK) and co-transfected into 293T cells with an *env* deficient backbone, nl4.3Δenv (Dr M. Pizzato, University of Geneva). Virus-containing supernatants were

harvested, assayed for reverse transcriptase activity (Pizzato et al. 2009), and titrated onto the HIV permissible cell-line, TZM-BL (also known as JC53-BL) using previously described techniques(Montefiori 2005) with the following modifications: cell monolayers were fixed with 0.2% gluteraldehyde, stained with an X-gal substrate and air dried. Infected cells were counted with an AID v2.9 ELISpot plate-counter (AID GmbH, Germany). To test serum-mediated neutralizing responses, 400 focus forming units (FFUs) of titrated-pseudovirus were incubated with serial dilutions of heat inactivated autologous sera from participants. Neutralization was calculated as the percentage-reduction of FFUs compared to virus-only controls.

3.3.4.3 IFN-γ ELISpot assay

100μl of 0.5μg/ml mouse anti-human IFN-γ monoclonal antibody solution (Mabtech, Sweden) was added to each well on an ELISpot plate (Millipore, US). Frozen PBMCs were rapidly defrosted and then pipetted into 10ml of a solution containing RPMI 1640 and pigskin gelatine (PSG) with added DNAse (Sigma, UK). The solution was centrifuged at 300 x *g* for 5 min. The PBMCs were re-suspended in 20 ml of R10 solution and incubated overnight at 37°C. Cells were then counted and re-suspended in a volume of R10 solution to give a final concentration of 5 x 10^5 cells per 100μl. The ELISpot plate was washed three times with 200μl per well of phosphate buffered solution (PBS; Gibco, US) containing 1% FCS. Peptides were added to the appropriated wells, with a final concentration of each peptide being 10μM. We used overlapping 15mer peptides covering HIV-1 proteins gag p17 and gag p24 as well as optimal epitopes covering gag, pol, nef and env proteins. 100μl of PBMC suspension was then added to each well. Duplicate negative controls were prepared,

containing R10. Duplicate positive controls were prepared, containing 5μg/ml PHA-P (Sigma, UK). The plate was incubated for 16 hours at 37°C. The PBMCs were then discarded and the plate was then washed seven times with PBS. 100μl of 0.5μg/ml biotinylated anti-human IFN-γ monoclonal antibody (Mabtech, Sweden) was added to each well. The plate was incubated for 90 min at room temperature. The antibody was then discarded and the plate washed seven times with PBS. 100μl of 0.5μg/ml streptavidin-conjugated alkaline phosphatase (ALP; Mabtech, Sweden) was added. The plate was incubated at room temperature for 40 min. The streptavidin-ALP was then discarded and the plate washed seven times with PBS. 100μl of substrate solution from the ALP conjugate substrate kit (Bio-Rad, US) was added to each well. The plate was incubated at room temperature for 10 min, or until a colour change was noted in the positive control well. The plate was then washed with ordinary tap water and dried. Spots were counted on the AID version 2.9 ELISpot plate-reader. The normalized magnitude of the response (*NMOR*) was calculated as follows (Duda et al. 2009):

$$NMOR = M_{\exp} - (\bar{x}_{neg} + 3 \times SD_{neg}) - 50$$

where $M_{\exp}$ is the number of spots in the experimental well, $\bar{x}_{neg}$ is the mean number of spots in the negative control wells, and SD_{neg} is the standard deviation of the negative control wells. *NMOR* is always a positive integer and all negative values are set to 0.

3.4 Results

3.4.1 Case history of a single, third party exposure and recent seroconversion

3.4.1.1 Clinical History

Two adult males, P1 and P2, reported a single sexual encounter each with the same third-party that occurred on day 0. P1 and P2 reported subsequent exposure only to each other prior to enrolment in the SPARTAC trial. Despite repeated efforts, the third-party donor could not be traced. On day 6 post-exposure, P1 presented to his primary care physician with symptoms compatible with HIV seroconversion.

3.4.1.2 Clinical laboratory data consistent with recent infection

Clinical and laboratory evidence supported recent seroconversion in P1 and P2. On day 25, P1 tested positive for HIV-1 by ELISA with an incident result on a detuned ELISA, suggestive of recent infection (Highleyman 1999; Kothe et al. 2003). P2 had a positive HIV-1 PCR and negative HIV-1 ELISA on day 22, and on day 35 was p24 positive but negative by Murex ELISA (R&D Systems, UK) (Fiebig et al. 2003). The Murex ELISA was repeated on day 56 and had become clearly positive. Although, the Murex ELISA was positive in P1 earlier than in P2, the result was consistent with reported between-host variability in both the duration of the pre-viraemic phase and the timing of the appearance of markers of seroconversion (Fiebig et al. 2003; Fiebig et al. 2005).

3.4.1.3 Evidence of exposure of P1 to a fourth host after baseline

P1 and P2 were sampled on the same day when they enrolled in the SPARTAC trial, 63 days post-exposure. Both participants were randomized to receive no therapy. Plasma for sequencing was re-

sampled on the same date from both participants, on day 235 post-exposure. P1 reported exposure to a fourth party after day 63 and before day 235. Evidence of HIV-1 super-infection in P1 was seen on plasma collected at day 235 (Chapter 5). On day 245, P1 was diagnosed with acute hepatitis C virus (HCV) infection (Figure 3.1 a. and b.) having been negative for HCV by PCR and antibody on day 29. Therefore all time-points after day 63 were excluded from further phylogenetic analysis in Chapter 3.

3.4.1.4 The CD4+ count and plasma viral load values

The CD4+ count and plasma viral load values for P1 and P2 are shown in Figure 3.1 a. and b.. Despite the same exposure, P1 and P2 followed different clinical courses. P1 maintained a CD4+ T cell count greater than 350 cells/mm^3 during the first 310 days of untreated infection compared with P2, who had only two CD4+ readings greater than 350 cells/mm^3 over the first 249 days of infection. The plasma viral load for P1 was consistently lower than that for P2 after day 96, with the exception of a second peak reading in P1 taken on day 249, after the detection of HIV-1 super-infection and acute HCV infection. Therefore, P2 appeared to progress more rapidly than P1.

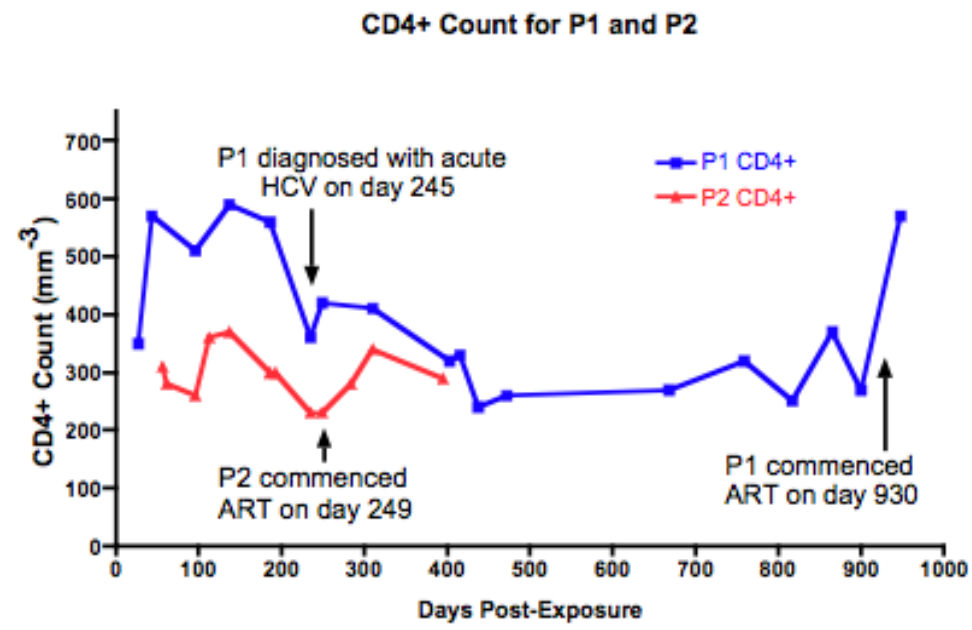


Figure 3.1 a. CD4+ count for P1 versus P2 over time

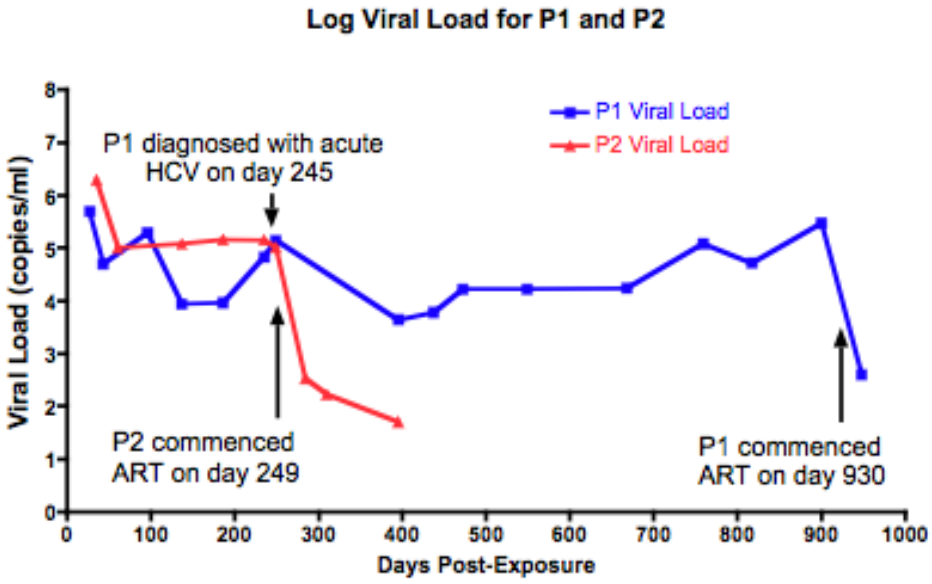


Figure 3.1 b. Log viral load for P1 versus P2 over time

3.4.1.5 Further clinical laboratory evidence was consistent with the history of a single donor

The time window for one participant to have infected the other was short. Participants P1 and P2 were both positive for p31 antigen on Western Blot on day 63. Therefore, the minimum estimated time since the onset of detectable viraemia (>50 copies/ml) of

approximately 47.4 days (Fiebig et al. 2003; Fiebig et al. 2005). Thus the estimated maximum pre-viraemic phase for either participant was 15 to 16 days. Since, the estimated pre-viraemic phase for HIV-1 lasts between 7 and 25 days (Clark et al. 1991; Fiebig et al. 2003; Fiebig et al. 2005; Little et al. 1999), one participant could have infected the other only between day 7 and day 9 post-exposure to the third party (Figure 3.1 c.).

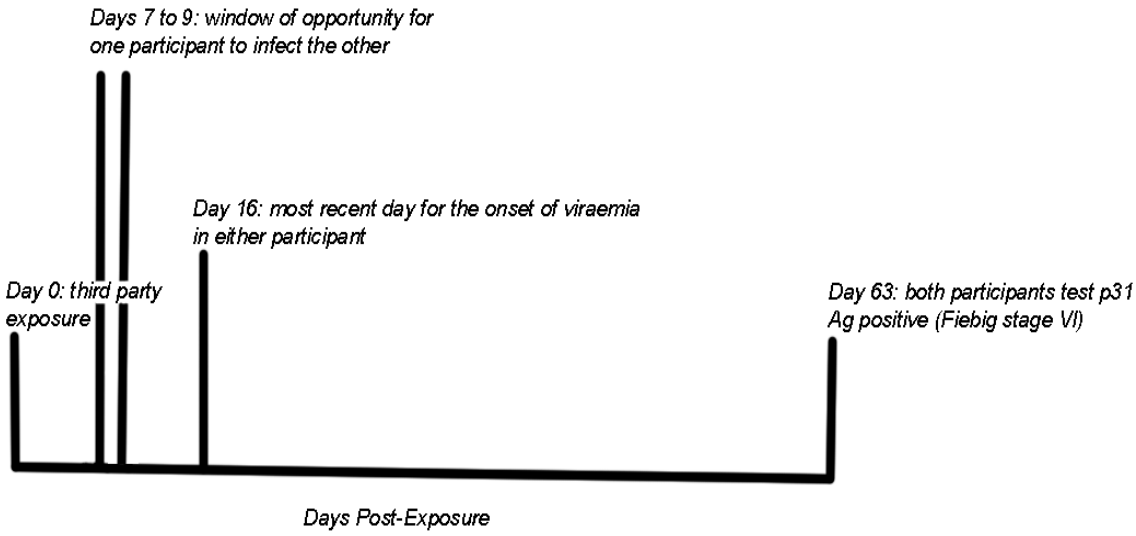


Figure 3.1 c. Key inferred events during the first 63 days of infection

However, peak viral load in acutely infected subjects is reached 7 or more days after the onset of detectable viraemia (Keele et al. 2008; Li et al. 2010; Stafford et al. 2000) and the infectiousness of a donor MSM is low if his viral load is 400 copies/ml of less (D. P. Wilson et al. 2008). Therefore, while the laboratory evidence did not exclude this alternative scenario, it was unlikely that one participant infected the other.

3.4.2 Sequences for phylogenetic analysis obtained from multiple viral genes

3.4.2.1 Summary of sampled genes or fragments

If P1 and P2 had indeed been infected by the same third person on the same night, I expected that viral sequences sampled from one recipient would be highly similar to sequences sampled from the other recipient. I sampled fragments of four different HIV-1 genes, 63 days post-exposure (Figure 3.2). The genes or fragments were located within the *env*, *nef*, *gag* and *pol* genes. The fragments are summarized in Figure 3.2.

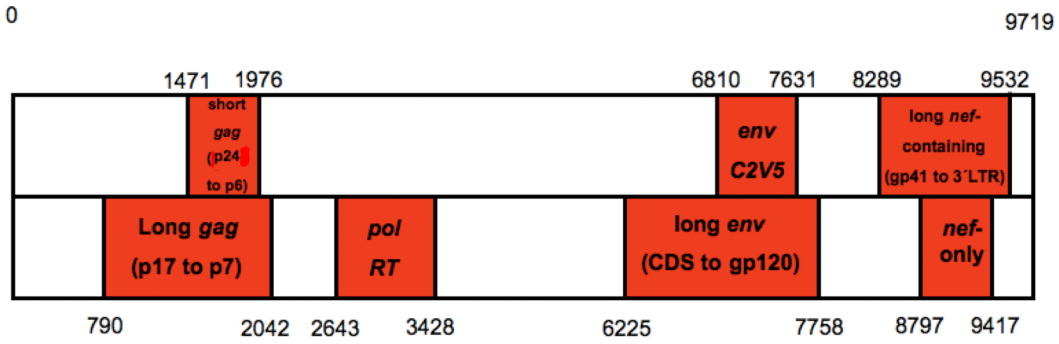


Figure 3.2 Sampled genes or fragments (not drawn to scale)

3.4.2.2 Env *fragments*

I sampled a long *env* fragment extending from the start of the *gp160* coding region to the end of the *gp120* coding region (HXB2 nucleotide position 6225 to 7758) by SGA (Bar et al. 2010; Keele et al. 2008; Li et al. 2010; Palmer et al. 2005; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009). After gap-stripping, the *env* gene fragment alignment was 1305bp in length. I also analysed a sub-section of this fragment, *env C2V5* (HXB2 position 6810 to 7631), to facilitate the molecular clock analysis in Chapter 6 (Lemey et al. 2006).

3.4.2.3 Gag *fragments*

A short *gag* fragment (*p24* to *p6*; HXB2 1471 to 1976) had been sampled by bacterial cloning (Jordan et al. 2010). I re-sampled a

longer *gag* fragment (*p17* to *p7*; HXB2 790 to 2042) by bacterial cloning (Jordan et al. 2010). The short *gag* fragment alignment was 449bp in length. The re-sampled long *gag* fragment alignment was 1250bp in length.

3.4.2.4 Nef *fragments*

I sampled a fragment extending from within *gp41* to within the *3´LTR* (HXB2 8289 to 9532) containing the nef gene (nef-containing fragment, Figure 3.2) by bacterial cloning. The *nef*-containing fragment alignment was 1239bp in length. Because this fragment overlapped with part of the *env gp41* gene, I also analysed a sub-region of the alignment containing only the *nef* gene (HXB2 8797 to 9417). The short *nef*-only gene alignment was 519bp in length.

3.4.2.5 Pol RT *fragment*

A fragment of *pol RT* (HXB2 2643 to 3428) was sampled by bacterial cloning (Jordan et al. 2010). The *pol RT* alignment was 724bp in length.

3.4.2.6 Reference sequences

I included reference sequences from individuals in the same geographical area and HIV-1 subtype. Reference sequences were drawn from the SPARTAC trial and the St Mary's Hospital Acute Infection Cohort (Fidler et al. 2002), as well as the LANL UK reference database (*gag*: N=52, *pol*: N=309, *env*: N=147, *nef*: N=70). Trees were rooted with outlier sequences from different HIV-1 subtypes and non-M groups in the LANL database. Sequences from both participants clustered with subtype B reference sequences in all of the fragments studied.

3.4.3 Between-host phylogenetic analysis is consistent with the clinical history of a single donor

By both ML and Bayesian MCMC based analyses, sequences from P1 and P2 were highly related and clustered to the exclusion of all other sequences, consistent with a common donor. I demonstrated statistical support for the robustness of the cluster by both methods. The resulting trees including all reference sequences were too large to interpret here so zoomed-in views of the trees have instead (Figures 3.3, 3.9, 3.13 and 3.19). By ML, the bootstrap values were: *env* 100%, *nef* 100%, *gag* 99.9% and *pol* 99.3%. Bayesian MCMC based posterior probabilities were: 100% for all of the fragments studied).

3.4.4 Phylogenetic signal-to-noise analysis

3.4.4.1 Phylogenetic signal-to-noise (STN) analysis in env

I quantified the percentage phylogenetic STN (Hillis and Huelsenbeck 1992) in *env*. I compared my full *env* fragment with gaps to the same fragment with GapStreeze set to 5%. The percentage STN between P1 and P2 was 70.7% to 24.3% in the unstripped *env* fragment and 62.0% to 30.7% for the stripped *env* alignment. Nevertheless, the percentage STN in the stripped alignment between recipient hosts was greater than in previous studies of multiple-variant transmissions in this genomic region (Keele et al. 2008; Li et al. 2010). Stripping the alignment of P1 and P2 to the shorter *env C2V5* region analysed Lemey and colleagues (Lemey et al. 2006) decreased the percentage signal (46.4%) and increased the percentage noise (53.2%) relative to the long *env* alignment. Noise ≥30% was consistent with a phylogenetic cluster (Lemey et al. 2009; Strimmer and von Haeseler 1997). My analyses indicated that there was a small loss of between-host phylogenetic signal in the long *env* fragment by stripping gaps or poorly aligned regions, and there was a much larger loss of signal with stripping to an *env C2V5* alignment. For the stripped alignments, noise greater than 30% was consistent with a phylogenetic cluster (Lemey et al. 2009; Strimmer and von Haeseler 1997).

3.4.4.2 Phylogenetic STN analysis in gag *and* pol

Stripped long env fragment alignments contained a higher percentage STN than either the short *gag* alignment (STN: 49.4% to 50.5%) or short *pol* alignment (STN: 64.2% to 35.5%). The *gag* and *pol* fragment between-host alignments did not require stripping. As with the *env* fragments, noise greater than 30% was consistent with a

phylogenetic cluster (Lemey et al. 2009; Strimmer and von Haeseler 1997).

3.4.4.3 Phylogenetic STN analysis in nef

The between-host alignments did not require gap-stripping for either the long *nef*-containing fragment or the short *nef*-only fragment. The percentage STN in both the long *nef*-containing (STN: 55.5% to 40.6%) and short *nef*-only (STN: 53.3 to 44.3%) fragments was larger than that in *env C2V5*, *pol* or *gag*, but not as great as the percentage STN in the long *env* fragment. As before, noise greater than 30% was consistent with a phylogenetic cluster (Lemey et al. 2009; Strimmer and von Haeseler 1997).

3.4.5 Phylogenetic analysis for sampled genes or fragments

3.4.5.1 Summary of findings

I compared the mean within-host distance, which is the average genetic distance between all the sequences within a host, to the between-host distance, which is the genetic distance along internal branches separating the within-patient sequences. With the exception of the larger *gag* fragment, the mean within-host distance was smaller than the between-host distance for all sampled fragments (Table 3.2). The lack of between-host diversity in *gag*,

Table 3.2 Summary of ML within- and between-host distance

gene (fragment)	Mean within-host distance (P1/P2)	Between-host distance	Fragment size (bp)
gag			
- *p24* to *p6*	0.26%/0.13%	0.54%	505
- *p17* to *p7*	0.40%/0.10%	0.24%	1252
env			
- long (cut)	0.54%/0.54%	3.81%	1305
- *C2V5* (cut)	0.25%/0.00%	5.73%	522
nef			
- *nef*-only	0.22%/0.23%	3.45%	519
- long	0.06%/0.24%	2.97%	1239
pol			
- *RT*	0.42%/0.22%	1.93%	724

relative to the mean within-host distance, suggested that P1 and P2 were infected with highly similar *gag* variants. This would be consistent with epidemiological linkage. The greatest differences in within-and between-host distance were seen in the *env* and *nef* fragments, suggesting the transmission of more divergent variants to each host. *Pol RT* yielded an intermediate result between *gag* and *nef*/*env*, with a high-than expected mean within-host distance in P1, when compared to a less-conserved region of the genome, such as *nef*. I tested the significance of these observations by conducting a coalescent analysis in Chapter 4.

3.4.5.2 Gag *fragments analysis*

Sequences from P1 and P2 clustered together to the exclusion of all other sequences in both *gag* fragments, with bootstrap and posterior probability supports of at least 99.9% and 100% respectively (Figure 3.3). These phylogenetic trees estimated by two

different methods were consistent with the *gag* sequences from P1 and P2 being epidemiologically linked. In particular, the long *gag* fragment sequences from the participants inter-mingled (Figure 3.3 c. and d.). The within- and between-host distance by ML estimation are shown in Table 3.2. However, the null hypothesis, in which the sequences were not epidemiologically linked, could not be excluded. Therefore, I used additional methods for comparing and contrasting the sequences from P1 and P2. One way to compare a relatively small set of sequences is to visually inspect the base substitutions, particularly for mutations that result in non-synonymous mutations. Non-synonymous mutations indicate changes to the amino acid sequence that may result from natural selection. Initially, I visually inspected the *gag* sequences within each participant for evidence of homogeneous infection. I looked for base substitutions in Highlighter plots of Day 63 long- and short- fragment *gag* sequences in P1 (Figures 3.4 and 3.5) and P2 (Figures 3.6 and 3.7). Highlighter plots also allowed for the capture of possible hypermutation signatures, which would not be captured by phylogenetic tree reconstruction lone, that may explain some of sequence evolution within each host. The sequences demonstrated frequent anti-sense (G to A) APOBEC hypermutation signatures (Figures 3.4 and 3.6). I concluded that the *gag* sequences were consistent with early

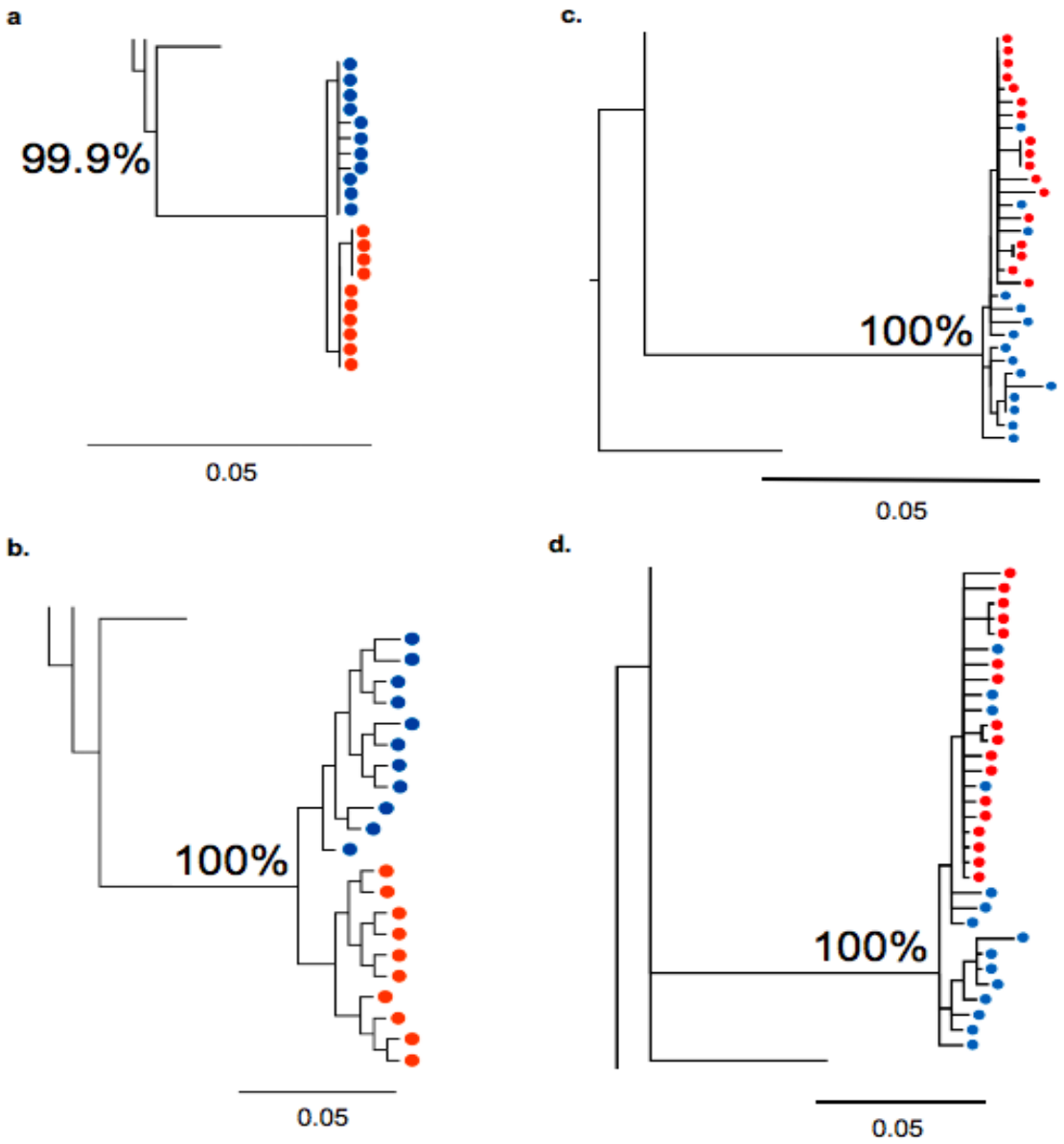


Figure 3.3 Zoomed-in phylogenetic trees of the small* gag *fragment (a. and b.) and the large* gag *fragment (c. and d.) for P1 (blue) and P2 (red). *ML trees (a. and c.) and Bayesian MCMC based trees (b. and d.) are shown. Bootstrap values (a. and c.) or posterior probabilities (b. and d.) for the cluster are also given as percentages. The scale bar shows genetic distance (0.05 = 5%).*

infection in each participant, although there was evidence of accumulation of some non-synonymous mutation relative to the consensus sequence.

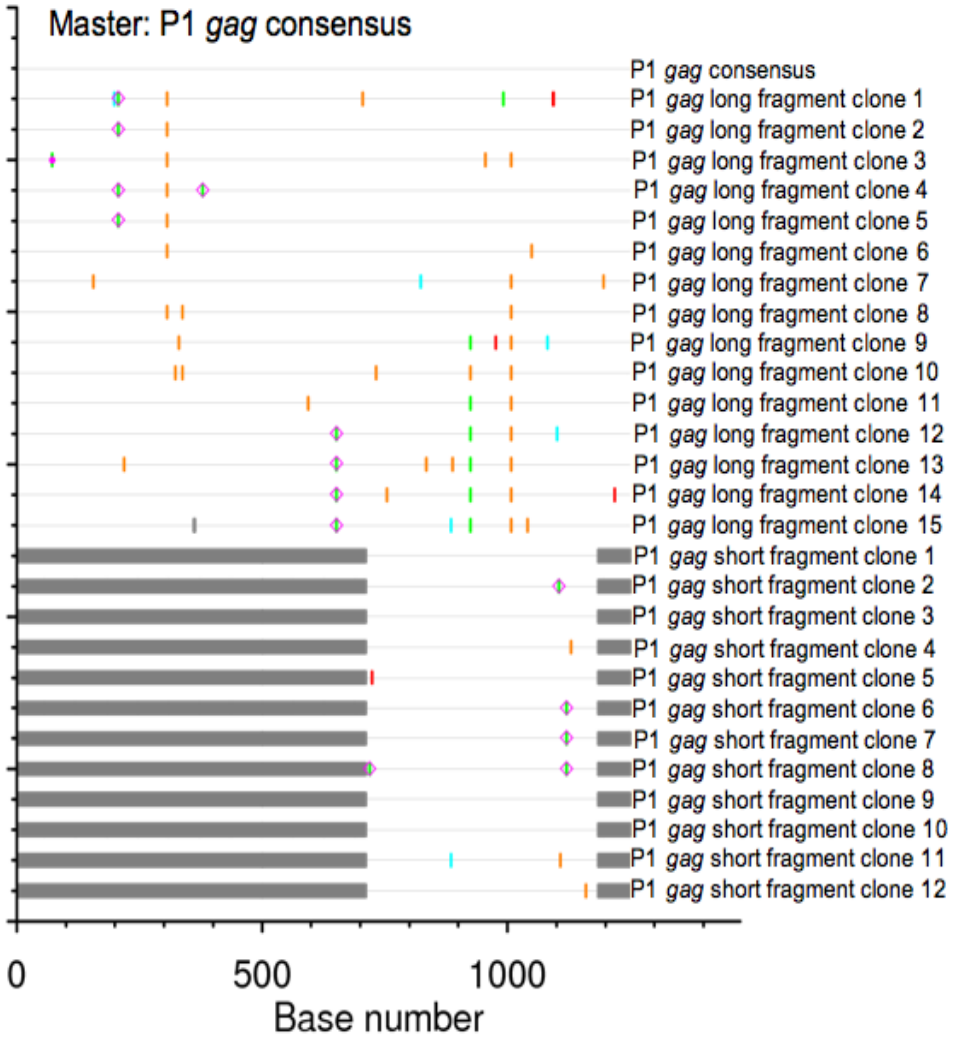
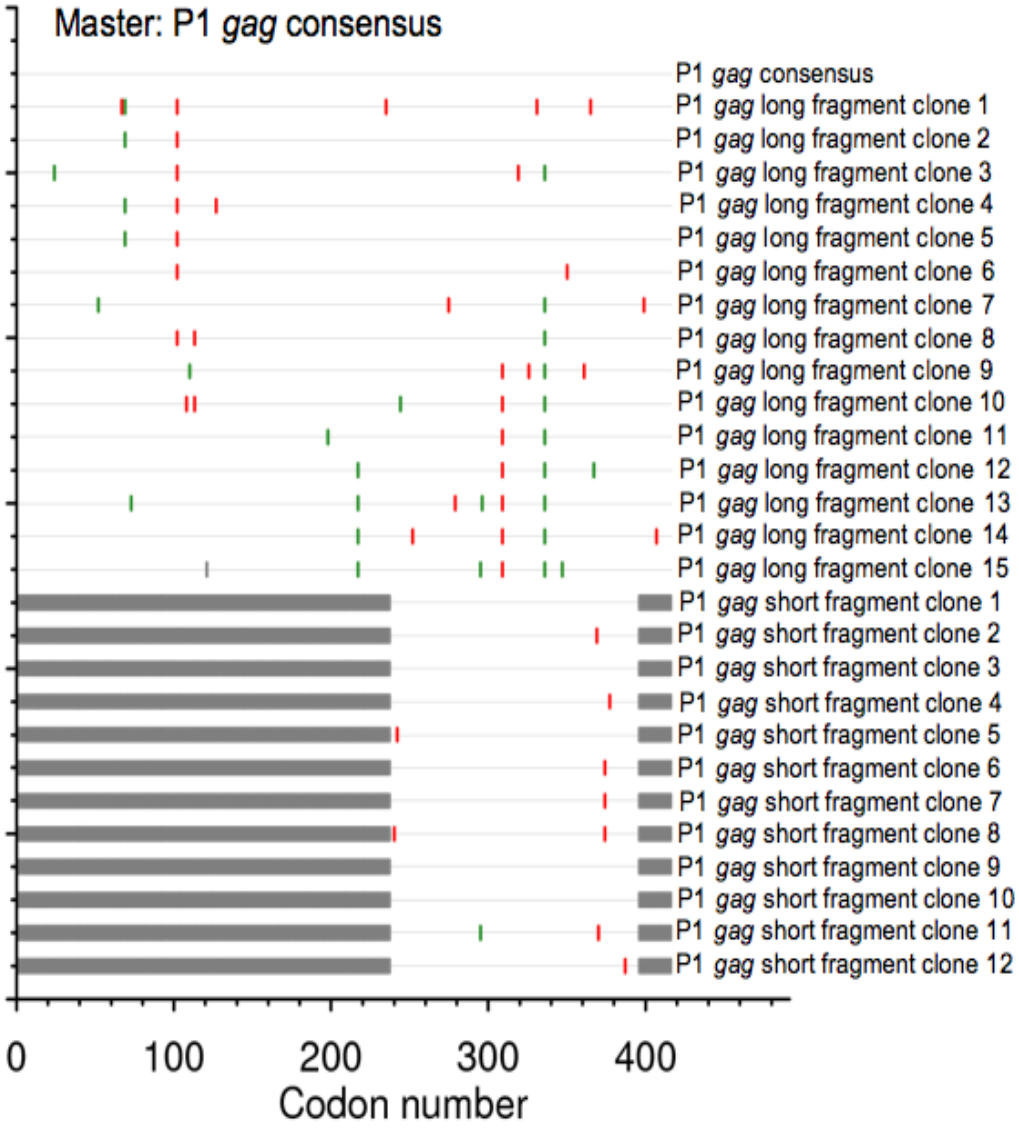


Figure 3.4 Highlighter plot of the* gag *fragment alignments for P1, showing base substitutions. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: A=green, T=red, G=orange, C=light blue, gap=grey, pink closed circle=APOBEC (A to G), pink open diamond=G to A (anti-sense APOBEC). Motifs consistent with APOBEC signatures are marked in purple.*



a

Figure 3.5 Highlighter plot of the* gag *fragment alignments for P1, showing non-synonymous versus synonymous mutations. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: synonymous=green, non-synonymous=red.*

Additionally, different variant populations were detected within the same host for the long- and short- fragments for both P1 and P2. This result may be attributable to primer-bias (Jordan et al. 2010) or

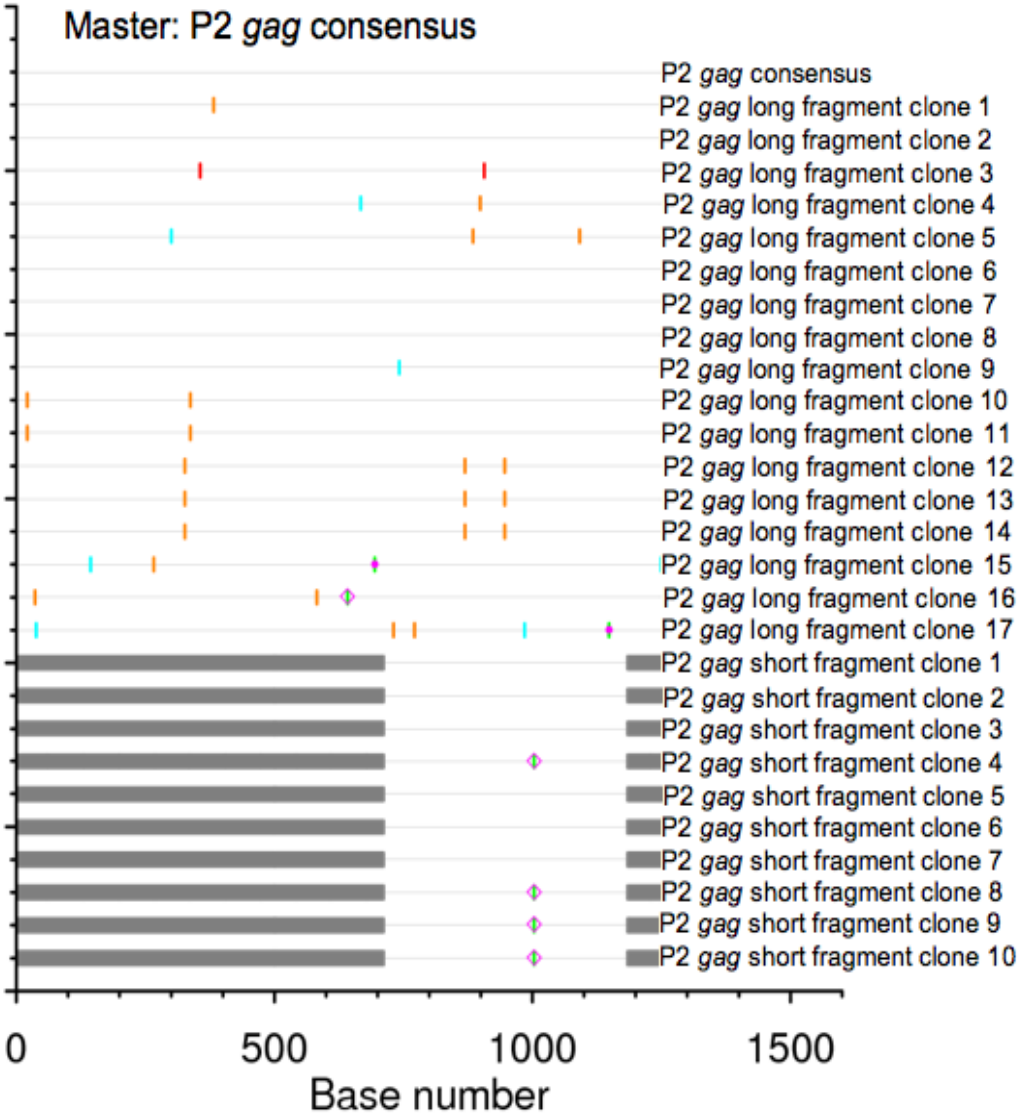


Figure 3.6 Highlighter plot of the* gag *fragment alignments for P2, showing base substitutions. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: A=green, T=red, G=orange, C=light blue, gap=grey, pink closed circle=APOBEC (A to G), pink open diamond=G to A (anti-sense APOBEC). Motifs consistent with APOBEC signatures are marked in purple.*

PCR-induced mutation, emphasising the importance of using multiple primers/PCR runs when amplifying very similar sequences.

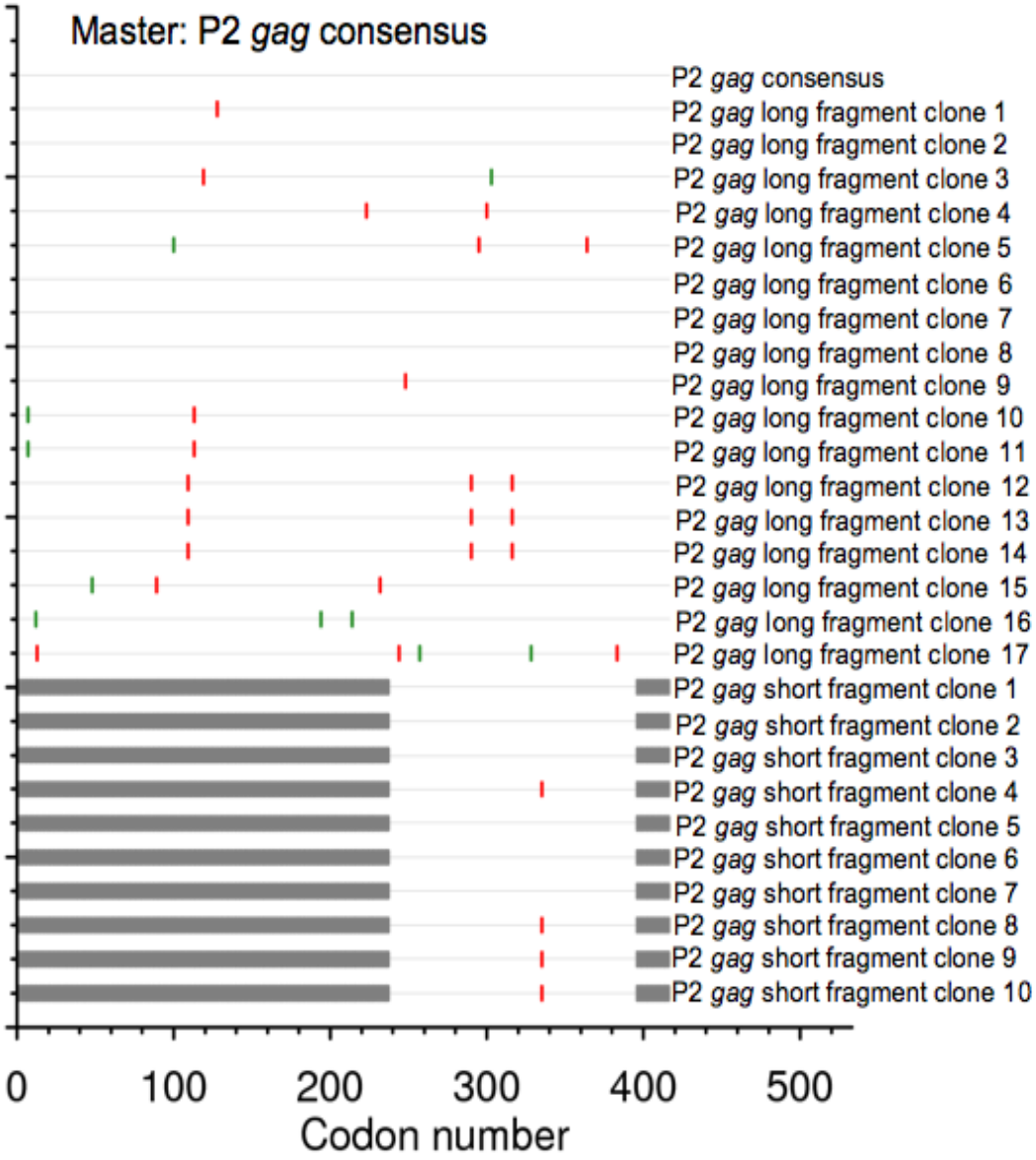


Figure 3.7 Highlighter plot of the* gag *fragment alignments for P1, showing non-synonymous versus synonymous mutations. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: synonymous=green, non-synonymous=red.*

A sub-population of seven long-fragment *gag* sequences in P1 (Figures 3.4 and 3.5) shared two or more substitutions with the long-fragment sequences in P2. The seven sequences also sat closest with the P2 sequences on both the ML and Bayesian MCMC based trees (Figure 3.8). Poisson Fitter rejected a star-like phylogeny for the long *gag* fragment in P1 but not P2. Poisson Fitter did not reject a star-

like phylogeny in the short fragment *gag* alignment in P1.

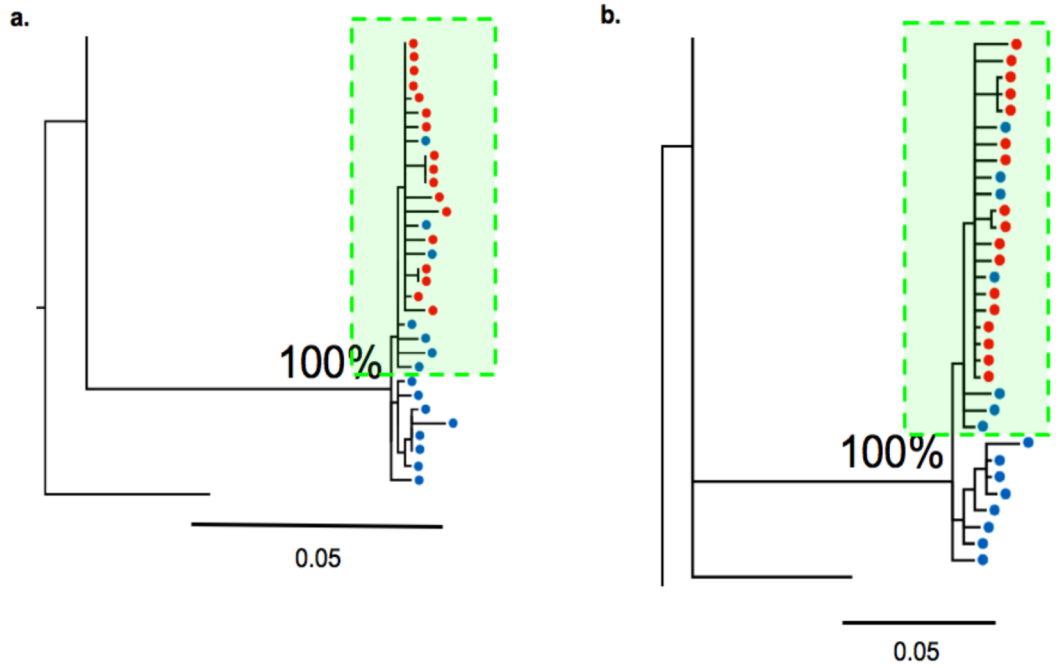


***Figure 3.8 Zoomed-in ML (a.) and Bayesian MCMC based (b.) trees for* gag.** *The regions of probable inter-mingling between P1 (blue) and P2 (red) sequences are highlighted in green dashed boxes. The scale bar shows genetic distance, so that 0.05 = 5%.*

The Hudson test rejected panmixia between P1 *gag* long-fragment clones 9 to 15 and the other P1 *gag* long fragment clones, with a p-value of 0.0004. The shared substitutions separating these seven sequences from the P1 consensus were a mixture of synonymous and non-synonymous mutations.

3.4.5.3 Env *fragments analysis*

The *long env* and *env C2V5* fragments were the least conserved fragments of those analysed, both within and between hosts. Sequences from P1 and P2 clustered together to the exclusion of all other sequences (Figure 3.9), with bootstrap and posterior probability support of 100%. The between-host distance for the *env* fragments are shown in Table 3.2.

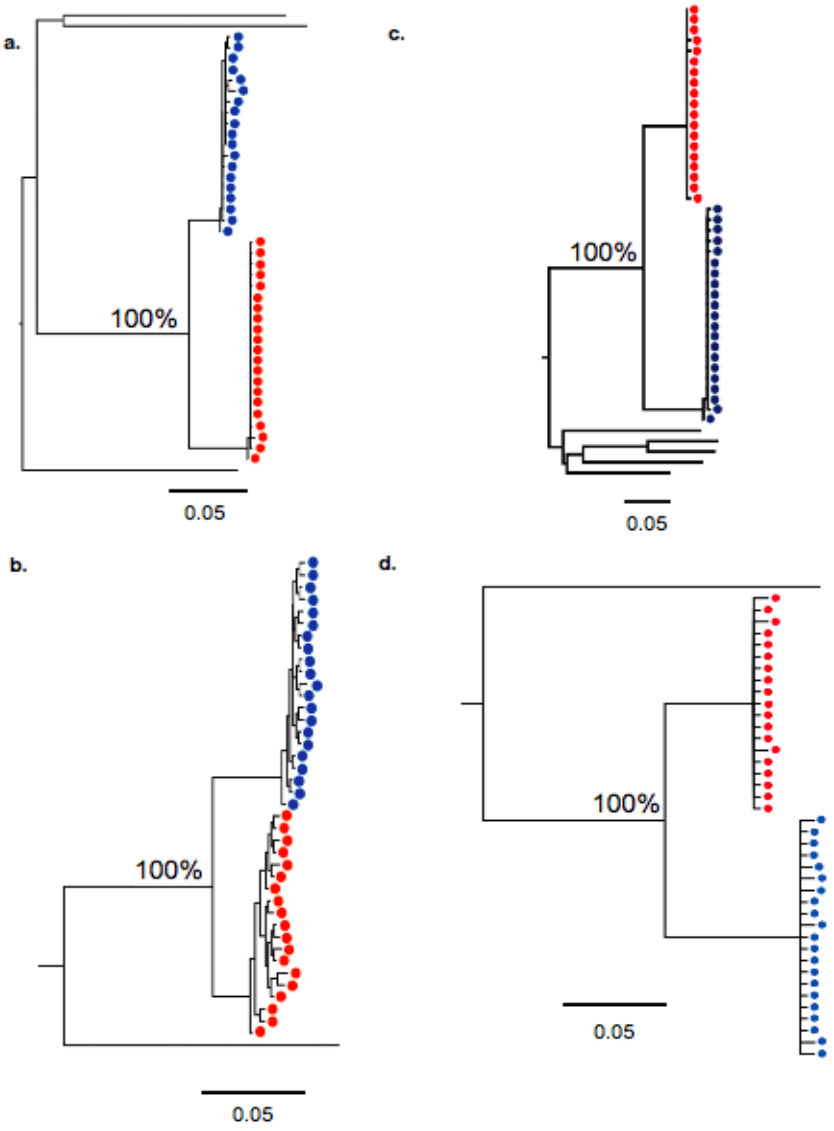


Figure 3.9 Zoomed-in phylogenetic trees of the large* env *fragment (a. and b.) and the* env C2V5 *fragment (c. and d.) for P1 (blue) and P2 (red). *ML trees (a. and c.) and Bayesian MCMC based trees (b. and d.) are shown. Bootstrap values (a. and c.) or posterior probabilities (b. and d.) for the cluster are also given as percentages. The scale bar shows genetic distance (0.05 = 5%).*

Phylogenetic analysis of the *env* region meant removing gaps and non-aligned portions, but current implementations of ML and Bayesian tree analysis do not model gaps or non-aligned regions informatively (Salemi et al. 2007). Therefore, I compared full-fragment, non-stripped *env gp120* sequences from P1 and P2 with the baseline consensus sequence for P1 in Highlighter (Figure 3.10).

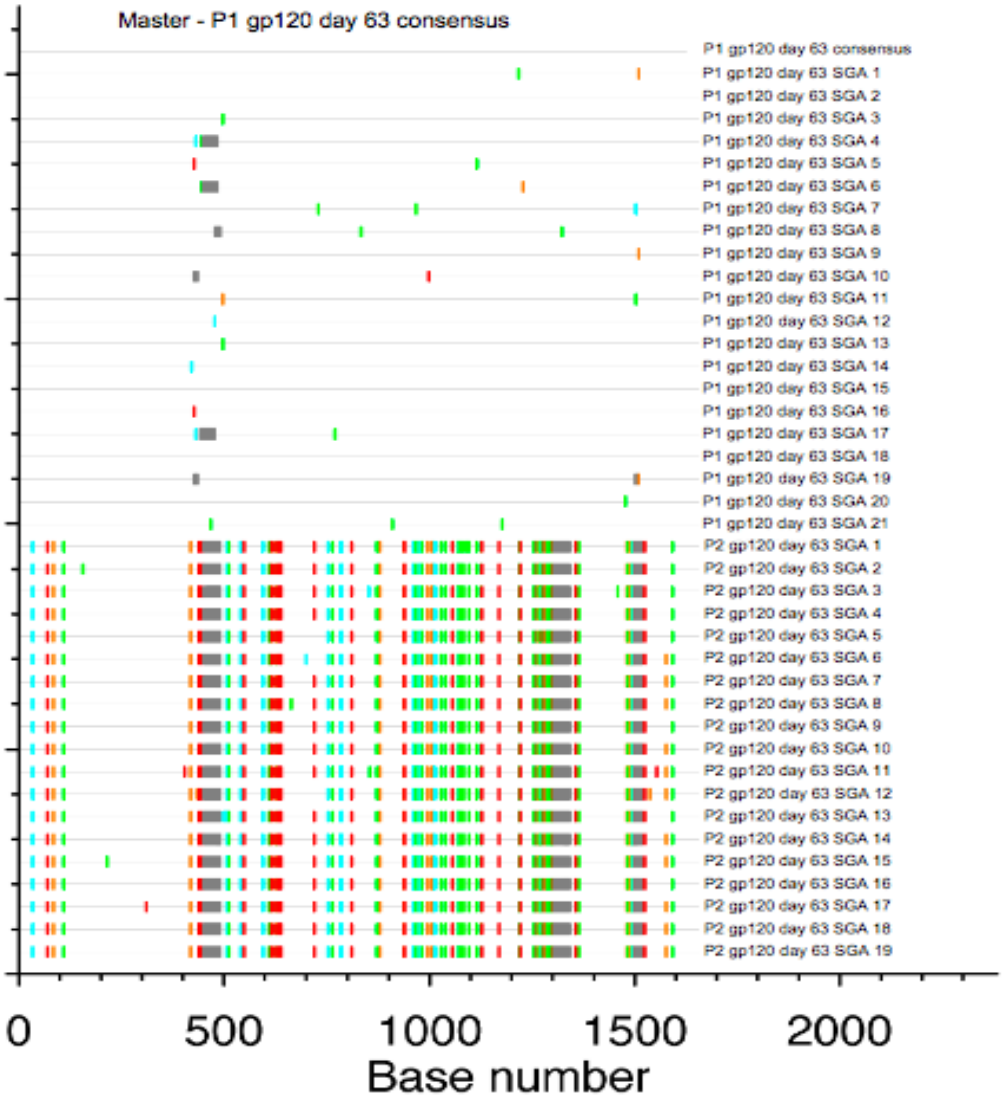


Figure 3.10 Highlighter plot of base mismatches in the unstripped* gp120 *region relative to the P1 gp120 day 63 SGA consensus. *The x-axis indicates base position along the sequence alignment. Sequence names are indicated on the y-axis. Colours are: A=green, T=red, G=orange, C=light blue, gap=grey.*

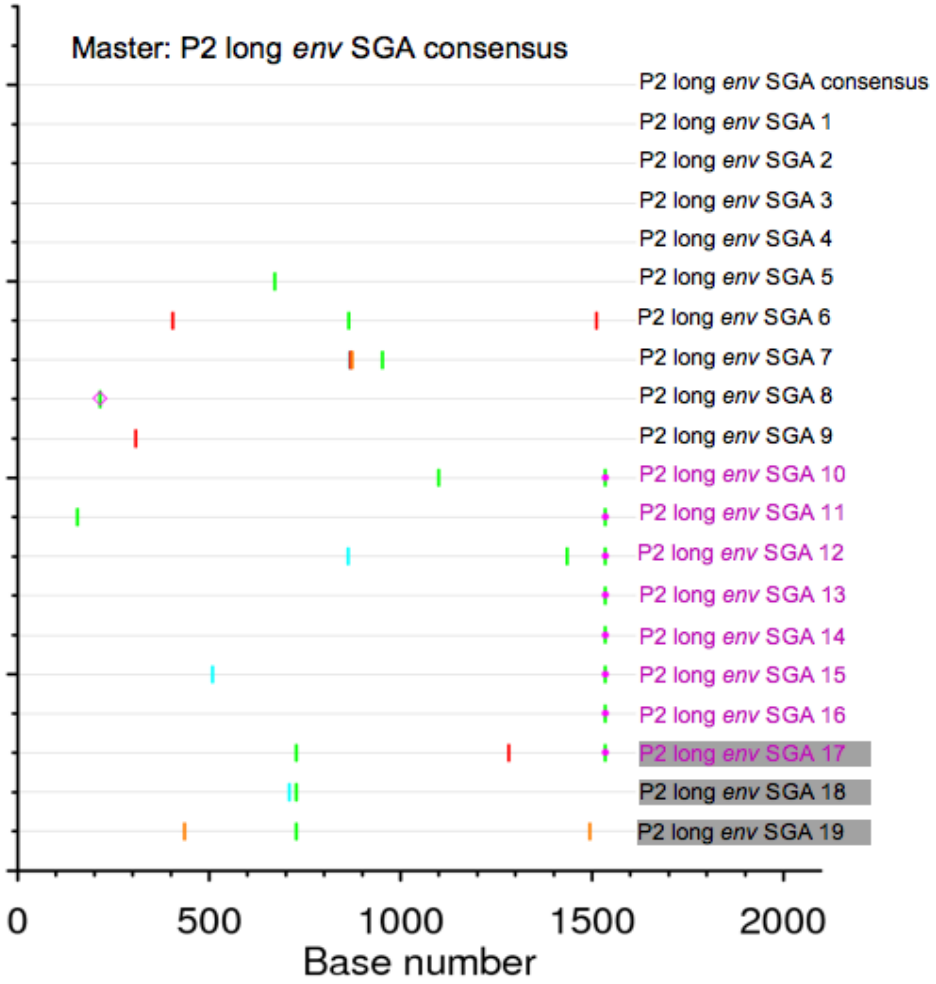


Figure 3.11 Highlighter plot of unstripped* gp120 *region showing base mismatches relative to the day 63 P2 SGA consensus. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: A=green, T=red, G=orange, C=light blue, gap=grey, pink closed circle=APOBEC (A to G), pink open diamond=G to A (anti-sense APOBEC). Sequences sharing a common G-to-A substitution are indicated by sequence names in purple font. Sequences sharing a common T-to-A substitution are indicated by sequence names in grey boxes.*

Poisson Fitter did not reject a star-like phylogeny in either *env* fragments P1 or P2 if APOBEC signatures were removed (Chapter 8). Poisson Fitter rejected a star-like phylogeny for P2 in the long *env* fragment sequences with APOBEC signatures included (Chapter 8).

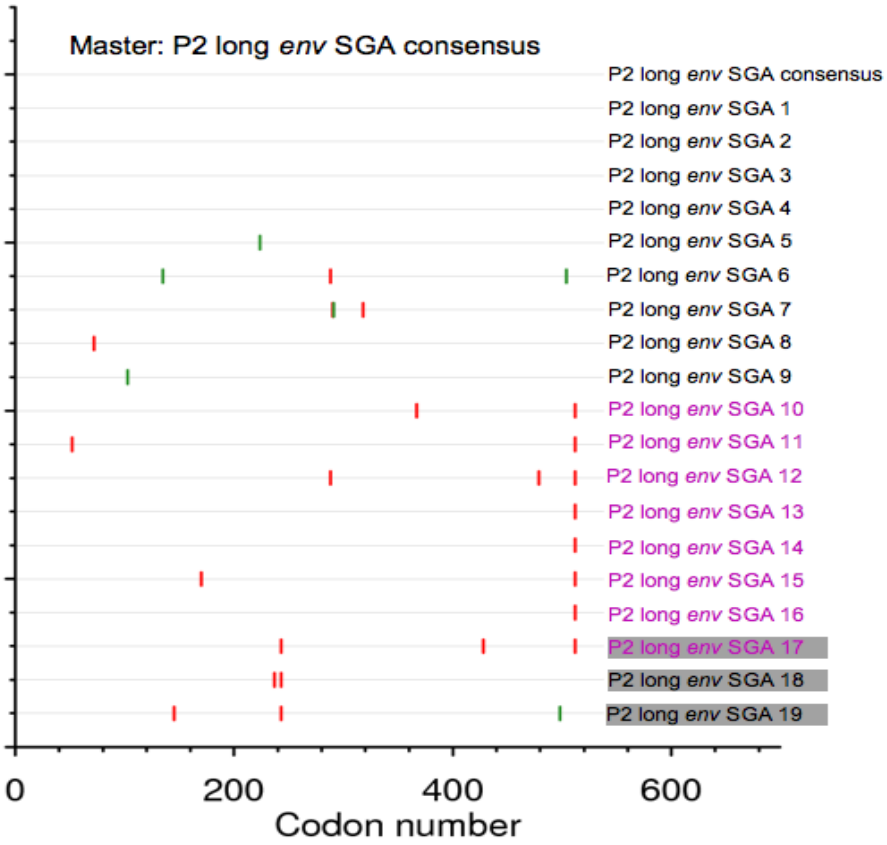


Figure 3.12 Highlighter plot of unstripped* gp120 *region showing synonymous (green) and non-synonymous (red) substitutions relative to the P2 day 63 SGA consensus. *The x-axis indicates codon position along the sequence alignment. The y-axis indicates sequence names. The colour key is given to the left of the left y-axis. Sequences sharing a common G-to-A non-synonymous substitution are indicated by sequence names in purple font. Sequences sharing a common T-to-A non-synonymous substitution are indicated by sequence names in grey boxes.*

The Hudson test rejected panmixia among the long-*env* fragment from P2 with a p-value of <0.0001. Poisson Fitter did not reject a star-like phylogeny in *env C2V5* with APOBEC signatures included (Chapter 8). Highlighter plots suggested two minor populations of sequences in P2 (Figures 3.11 and 3.12).

3.4.5.4 Nef *fragments analysis*

As with *env* and *gag*, the *nef* fragments clustered together to the exclusion of all other fragments (Figure 3.13). The between-host distances for each fragment are given in Table 3.2.

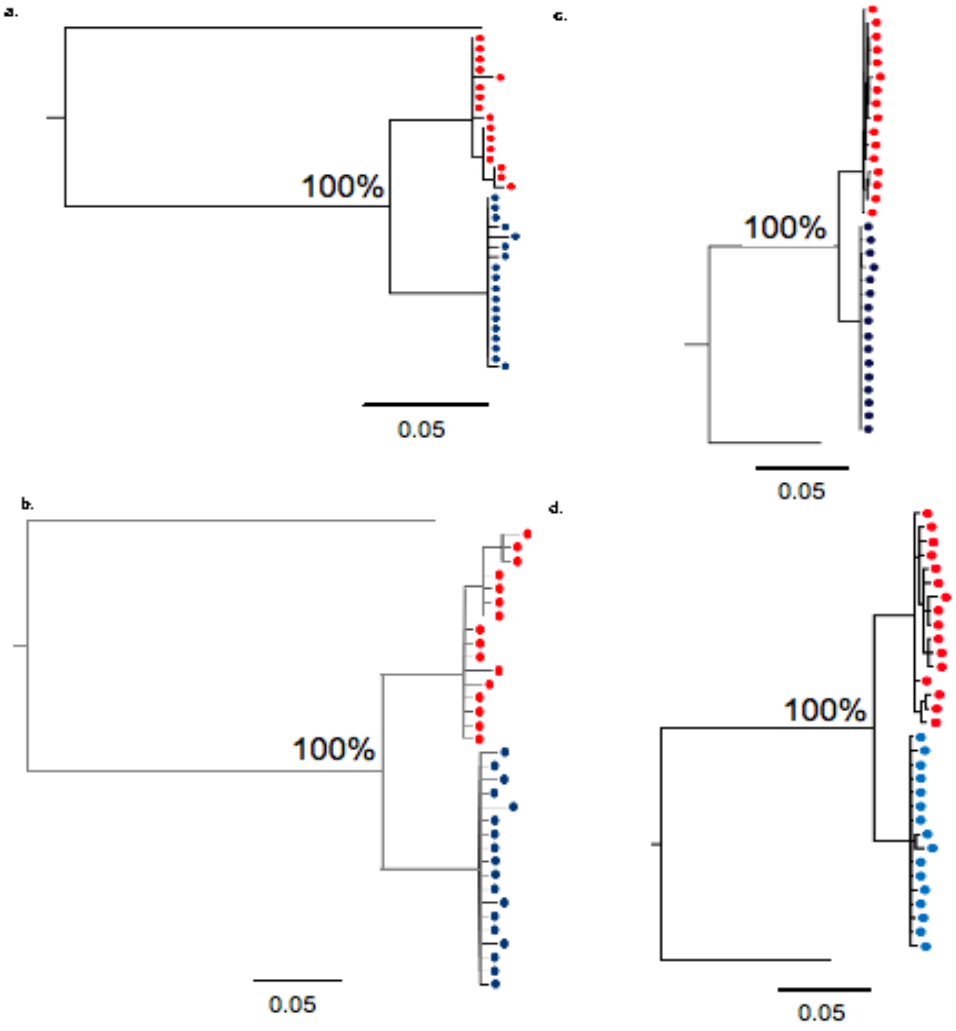


Figure 3.13 Zoomed-in phylogenetic trees of the* nef*-only fragment (a. and b.) and the large* nef*-containing fragment (c. and d.) for P1 (blue) and P2 (red). *ML trees (a. and c.) and Bayesian MCMC based trees (b. and d.) are shown. Bootstrap values (a. and c.) or posterior probabilities (b. and d.) for the cluster are given as percentages. The scale bar shows genetic distance (0.05 = 5%).*

The between-host distance in both *nef* fragments was less than for either *env* fragment but greater than that seen in the *pol* or *gag* fragments (Table 3.2). Poisson-Fitter did not reject a star-like phylogeny in either the long *nef*-containing fragment or the shorter *nef*-only fragments for P1 (data not shown). Poisson-Fitter rejected a star-like phylogeny in both the long *nef*-containing fragment or the shorter *nef*-only fragments for P2.

A Highlighter plot (LANL, US) of the *nef*-only fragment (Figures 3.14 and 3.15), revealed non-synonymous substitutions at between one and three shared codon positions in 8/16 sequences.

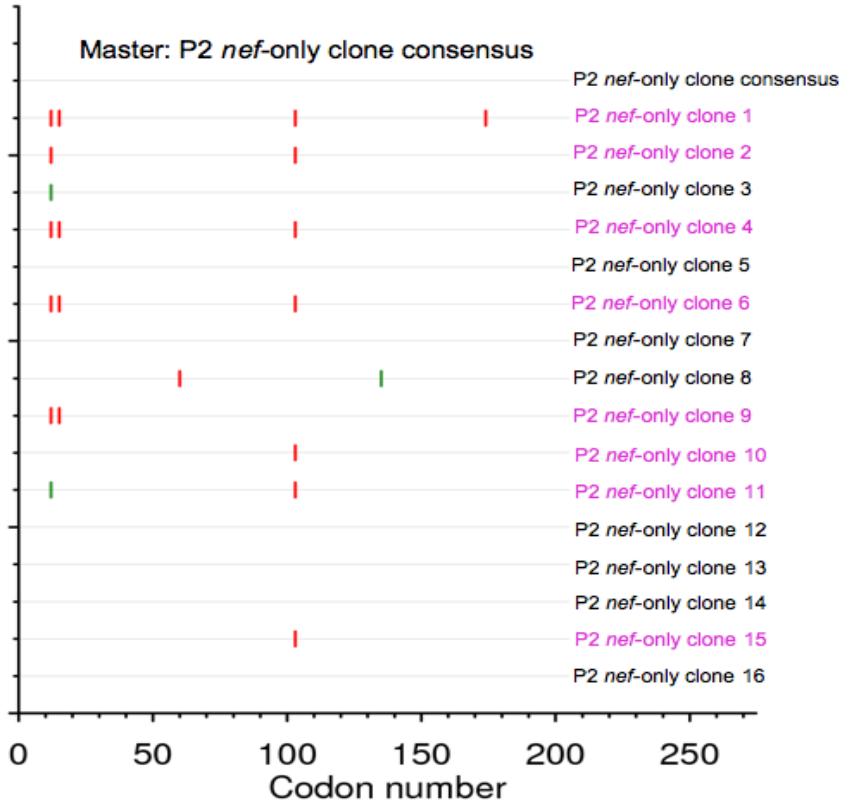


Figure 3.14 Highlighter plot of the* nef*-only fragment showing synonymous versus non-synonymous mutations relative to the P2 day 63 clonal sequence consensus. *Codon position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: synonymous=green, non-synonymous=red. Sequences with a non-synonymous mutation in at least one codon position in common are indicated by sequence names in purple font.*

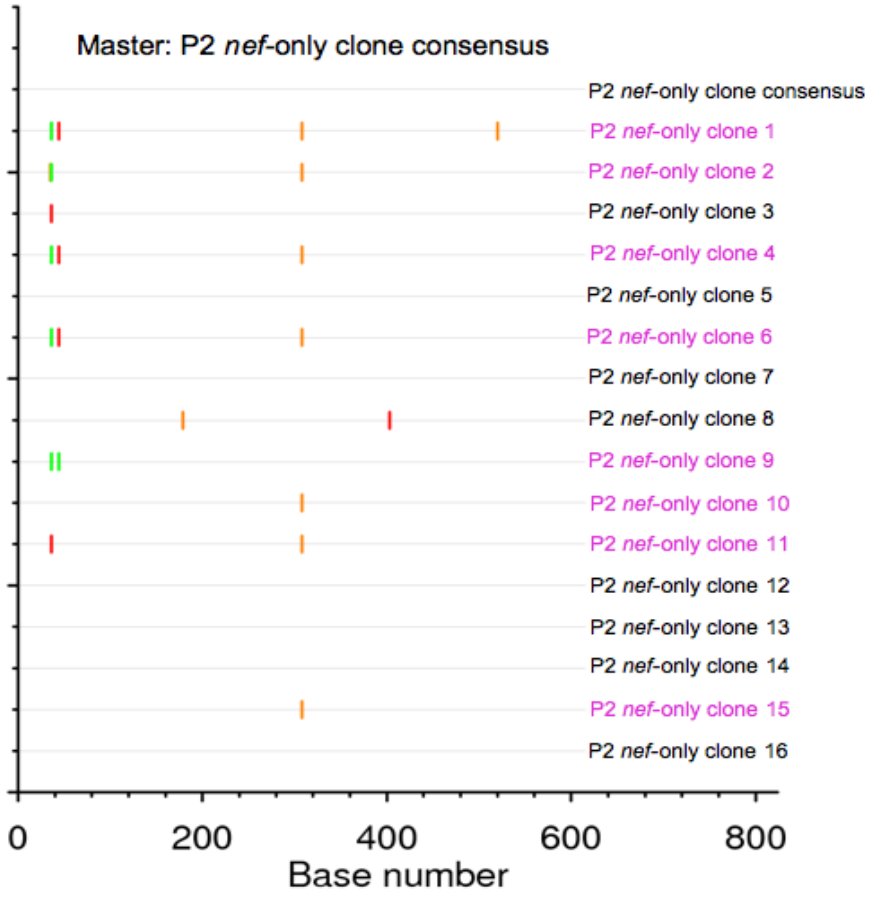


Figure 3.15 Highlighter plot of the* nef*-only fragment showing base mismatches relative to the P2 day 63 clonal sequence consensus. *Base position along the sequence alignment is indicated by the x-axis. Sequence names are indicated on the y-axis. Colours: synonymous=green, non-synonymous=red. Sequences with a nonsynonymous mutation in at one codon position in common are indicated by sequence names in purple font.*

The non-synonymous substitutions were at alignment codon positions 12, 15 and 103, corresponding to HXB2 *nef* codon positions 12, 15 and 105. The non-synonymous mutations at codon positions 12 and 15 were not homogeneous in terms of nucleotide substitutions (Table 3.3).

3.4.5.5 Pol *fragment analysis*

For *pol*, each recipient was infected with highly related sequences but multi-variant infections were detected in P1 and P2. The within- and between-host distance are given in Table 3.2. Poisson-fitter rejected a star-like phylogeny for P2 but not P1. Visual inspection of

the sequences with Highlighter was suggestive of multiple variant infection in P1 (Figure 3.16) but not P2 (Figure 3.17).

Table 3.3 Codon substitution patterns for HXB2, the P1 and P2 consensus sequences, and eight P2 clones at three codon positions. *Differences from the P2 consensus (**) are shown in bold and underlined. Codon position is numbered as the HXB2 position. Nucleotides and amino acids are named according to the single-letter code.*

Sequence name	Sequence: Nucleotide (amino acid)		
	Codon 12	Codon 15	Codon 105
HXB2	G**GA** (**G**)	**AC**T (**T**)	A**G**A (**R**)
Consensus P1	GA**A** (**E**)	GCT (A)	AAA (K)
Consensus P2	GAC (D)	GCT (A)	AAA (K)
P2 *nef* clone 01	GA**A** (E)	G**T**T (**V**)	A**G**A (**R**)
P2 *nef* clone 02	G**GA** (**G**)	GCT (A)	A**G**A (**R**)
P2 *nef* clone 04	GA**A** (**E**)	G**T**T (**V**)	A**G**A (**R**)
P2 *nef* clone 06	GA**A** (**E**)	G**T**T (**V**)	A**G**A (**R**)
P2 *nef* clone 09	GA**A** (**E**)	G**A**T (**D**)	AAA (K)
P2 *nef* clone 10	GAC (D)	GCT (A)	A**G**A (**R**)
P2 *nef* clone 11	GAC (D)	GCT (A)	A**G**A (**R**)
P2 *nef* clone 15	GAC (D)	GCT (A)	A**G**A (**R**)

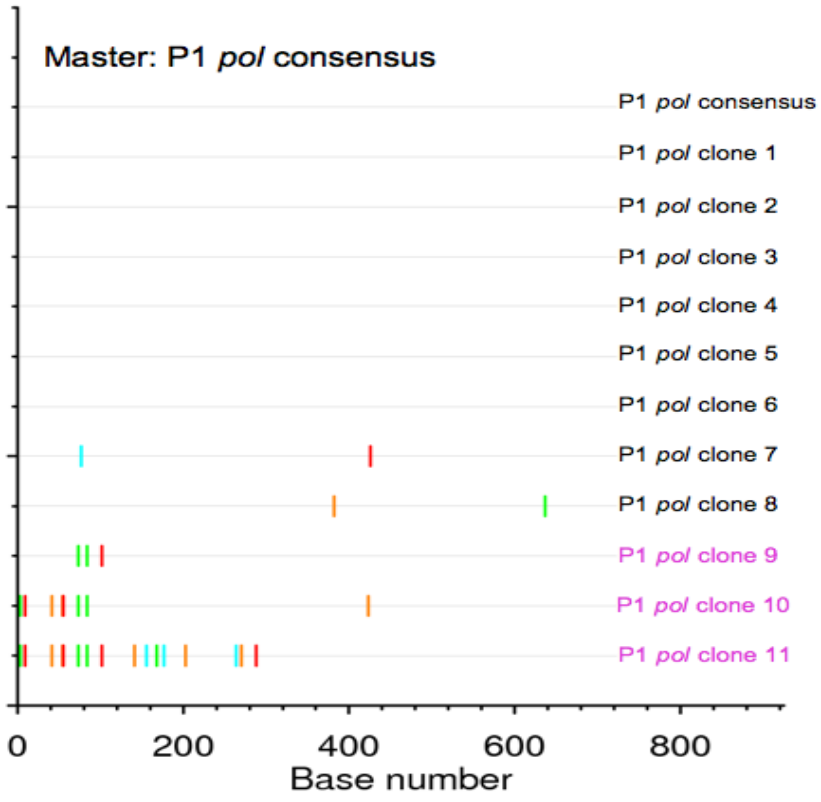


Figure 3.16 Highlighter plot for P1* pol *fragment base mismatches relative to the P1 day 63 clonal sequence consensus. *Base position along the sequence alignment is indicated on the x-axis. Sequence names are indicated on the y-axis. The names of clonal sequences representing suspected minor variants are indicated by purple text. Colours: A=green, T=red, G=orange, C=light blue, gap=grey.*

However, on inspection of the Highlighter plot for P2 (Figure 3.17), a sub-population of five clones were found to have non-synonymous mutations at codon position 239 (HXB2 *pol RT* codon position 274).

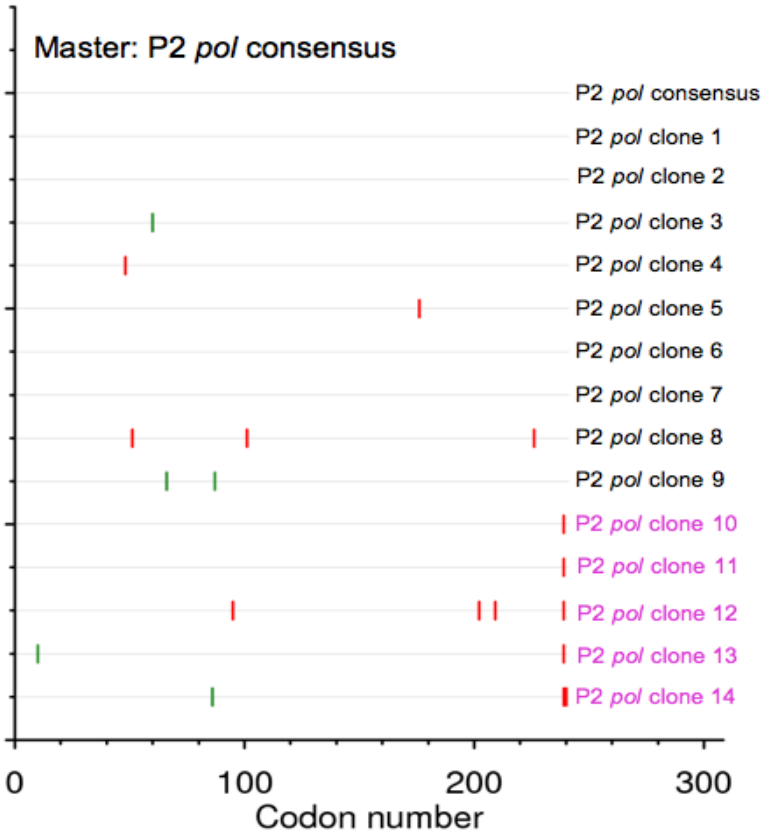


Figure 3.17 Highlighter plot for synonymous versus nonsynonymous substitutions in the P2* pol *fragment relative to the P2 day 63 clonal sequence consensus. *Codon position along the sequence alignment is indicated on the x-axis. Sequence names are indicated on the y-axis. The names of clonal sequences representing suspected minor variants are indicated by purple text. Colours: synonymous=green, non-synonymous=red.*

Although the five suspected minor variant sequences in P2 all had a non-synonymous amino acid substitution at the same codon

position, the nucleotide sequences were not homogeneous (Table 3.4). However, the Hudson test was performed to verify what was intuited from the Highlighter Plots. The Hudson test rejected population panmixia between the highlighted and non-highlighted

Table 3.4 Codon substitution patterns for HXB2, the consensus sequence for each recipient and the five suspected minor variant sequences in P2. *Differences from the P2 consensus (**) are shown in bold and underlined.*

Sequence	Codon sequence
HXB2	ATT
Consensus P1	ATT
Consensus P2	ATT
P2 pol clone 10	A**A**T
P2 pol clone 11	A**GA**
P2 pol clone 12	A**GG**
P2 pol clone 13	A**G**T
P2 pol clone 14	A**AG**

sequences in P1 (Figure 3.16) with a p-value of 0.0005. The Hudson test also rejected population panmixia between the highlighted and non-highlighted sequences in P2 (Figure 3.17) with a p-value of 0.0011. The suspected sub-populations of virus in P1 and P2 also clustered together on the phylogenetic trees and are highlighted in Figure 3.18, using a significance level of ≥ 70% (Figure 3.18). In particular, two variants from P1 clustered with a bootstrap of 94.3% and posterior probability of 97.0%, by ML and Bayesian analysis respectively (Figure 3.18). I also noted that tree-reconstruction suggested that P1 was likely to have more than one transmitted variant but Poisson Fitter had not rejected a star-like phylogeny for P1. This suggested a limitation to the utility of the algorithm of Poisson Fitter in detecting within-host subpopulations that might be suggested by visual inspection or phylogenetic tree reconstruction.

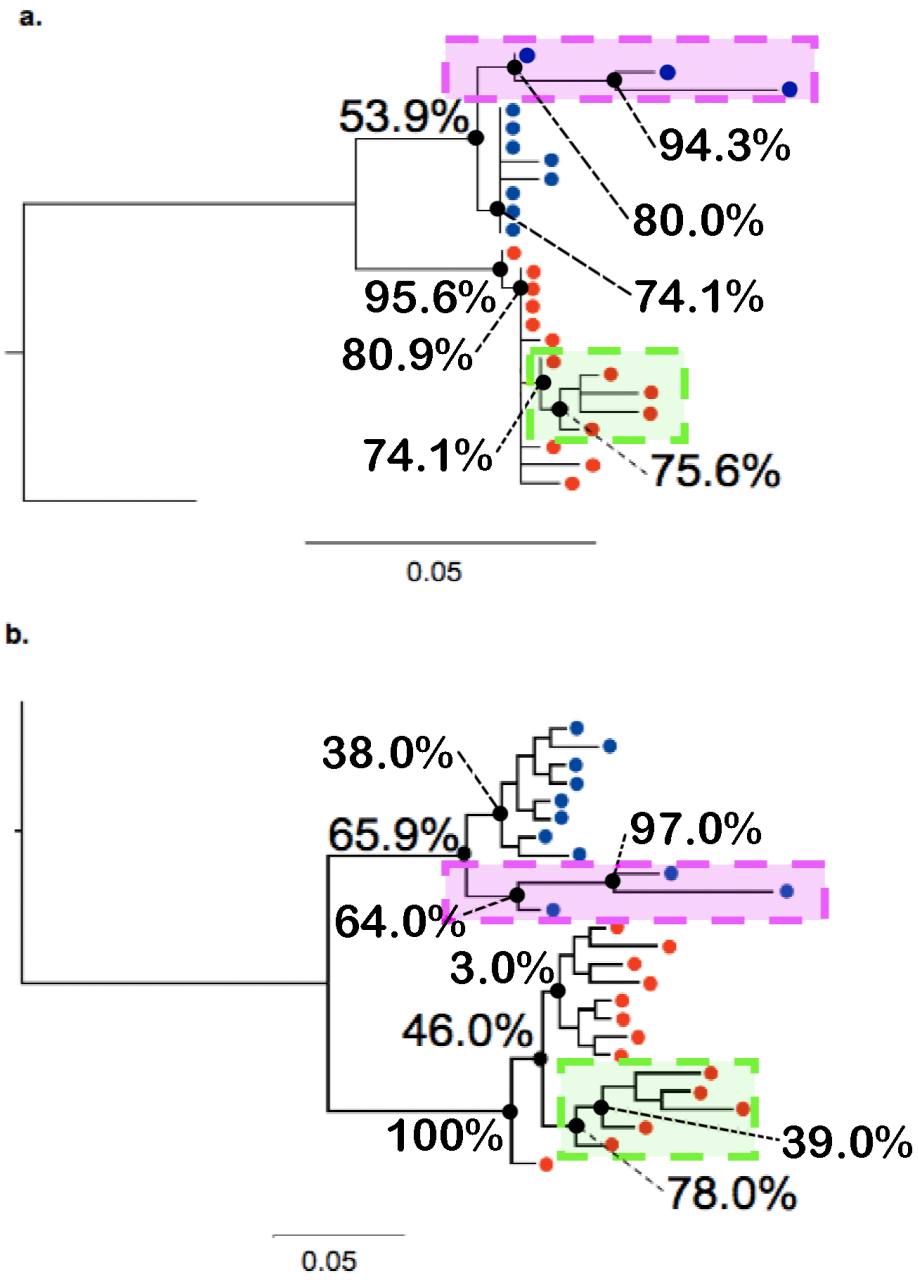


Figure 3.18 Zoomed-in ML (a.) and Bayesian MCMC (b.) tree for* pol *clonal sequences showing (a.) bootstrap or (b.) posterior probability values for internal branches (terminating at nodes indicated by black dots) for P1 (blue dots) and P2 (red dots). *Suspected minority variants highlighted in coloured dashed boxes for P1 (purple) and P2 (green). The scale bar shows genetic distance, so that 0.05 = 5%.*

3.4.6 Comparison transmission recipient pair (R1 and R2)

I attempted to validate my findings in P1 and P2 by investigating another suspected recipient pair. R1 and R2 were male participants in another cohort from the Royal Berkshire Hospital, Reading. R1

and R2 were also a couple who reported sharing a casual MSM exposure. They had been exposed to two high-risk contacts within the previous month. Each episode had involved both R1 and R2 being exposed to a single third party on the same evening. Despite extensive efforts, these contacts could not be traced. Therefore, in contrast to P1 and P2, there was no definite date of exposure to a single donor for R1 and R2. R1 had previously tested negative to HIV-1 on day -571 prior to presentation to a sexual health clinic (day 0). On day 0, R1 was positive for HIV-1 RNA and p24 Ag as well as weakly Ab positive. R2 presented to the sexual health clinic on day +20. On day +20, R2 was positive for HIV-1 RNA and p24 Ag, as well as strongly positive for HIV-1 Ab. Western Blot data were not available for R1 and R2. Unlike P1 and P2, sequences from R1 and R2 were sampled on different days and sequences were available from within the first four months of infection. The first available time-point in R1 was day +165 and the first available time point in R2 was day +207. Despite the high viral load in R2, *env* and *gag* sequences were sampled by bacterial cloning. Phylogenetic ML trees for R1 and R2 are shown for *env* and *gag* in Figure 3.19.

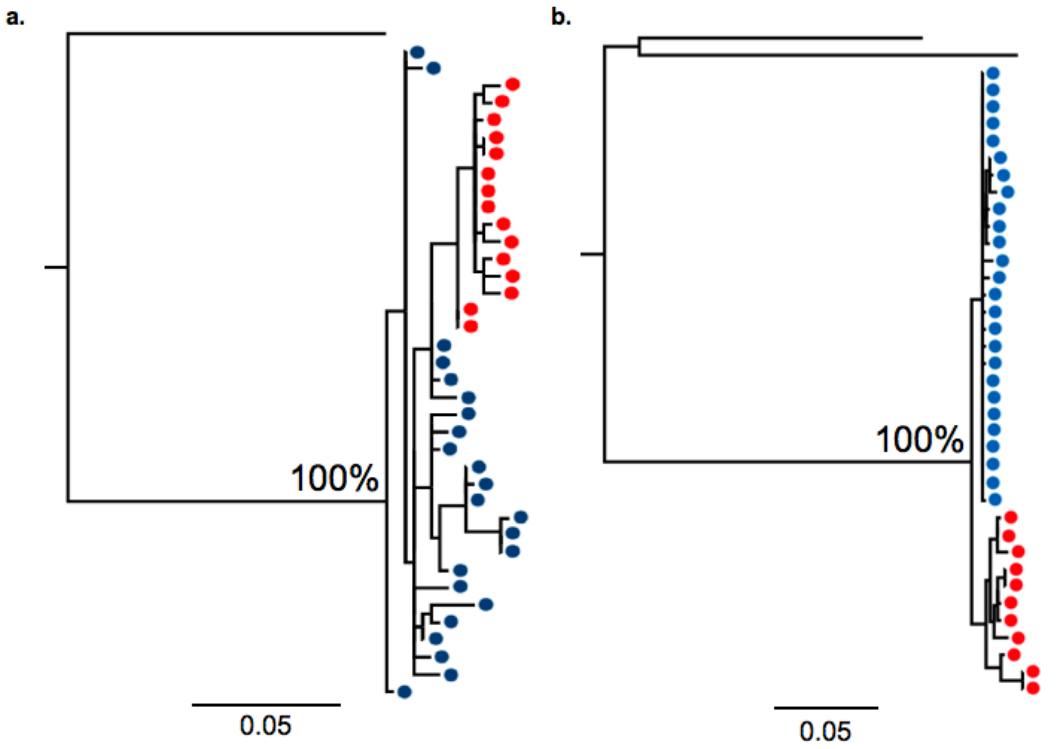


Figure 3.19 ML trees for a.* gag *and b.* env *in R1 (blue) and R2 (red). *Bootstrap support for the cluster is also given as a percentage. The scale bar shows genetic distance, so that 0.05 = 5%.*

As with P1 and P2, study baseline *env* and *gag* sequences for R1 and R2 clustered to the exclusion of all other sequences, with bootstrap support of 100%. This would be consistent with a single third party donor. As with P1 and P2, the *gag* sequences for R1 and R2 had a mean between-host distance of 0.99%. The mean within-host distances were 0.35% for R1 and 0.17% for R2. This was consistent with early chronic infection. However, in contrast to P1 and P2, the mean between-host distance in the *env* fragment was only 0.63%, which was less than that for the *gag* fragment. The low between-host divergence in *env*, coupled with uncertainty surrounding the date of exposure meant that it remained highly probable that R1 transmitted the infection to R2, or vice versa.

3.4.7 Differences in potential antigenic *env* variation between transmitted viruses in P1 and P2

3.4.7.1 Potential N-linked glycosylation sites (PNLGSs)

Demonstrating a high level of divergence did not answer whether or not P1 and P2 each received different *env* variants due to selection or to a random process. If transmission of divergent *env* variants was not random, then it may be caused by differences in antigenicity. *In vivo*, envelope proteins are under particular selection by neutralizing antibodies (Moore et al. 2009). The antigenicity of envelope proteins to neutralizing antibodies had previously been inferred from the number of PNLGSs in gp120 (Derdeyn et al. 2004; Simmonds et al. 1990). Therefore, I hypothesized that differences in the number of PNLGSs in gp120 would indicate potential between-host differences in viral antigenicity. I compared PNLGSs within inferred amino acid sequences for the gp120 protein from P1 and P2 using N-Glycosite (LANL, US; Figure 3.20).

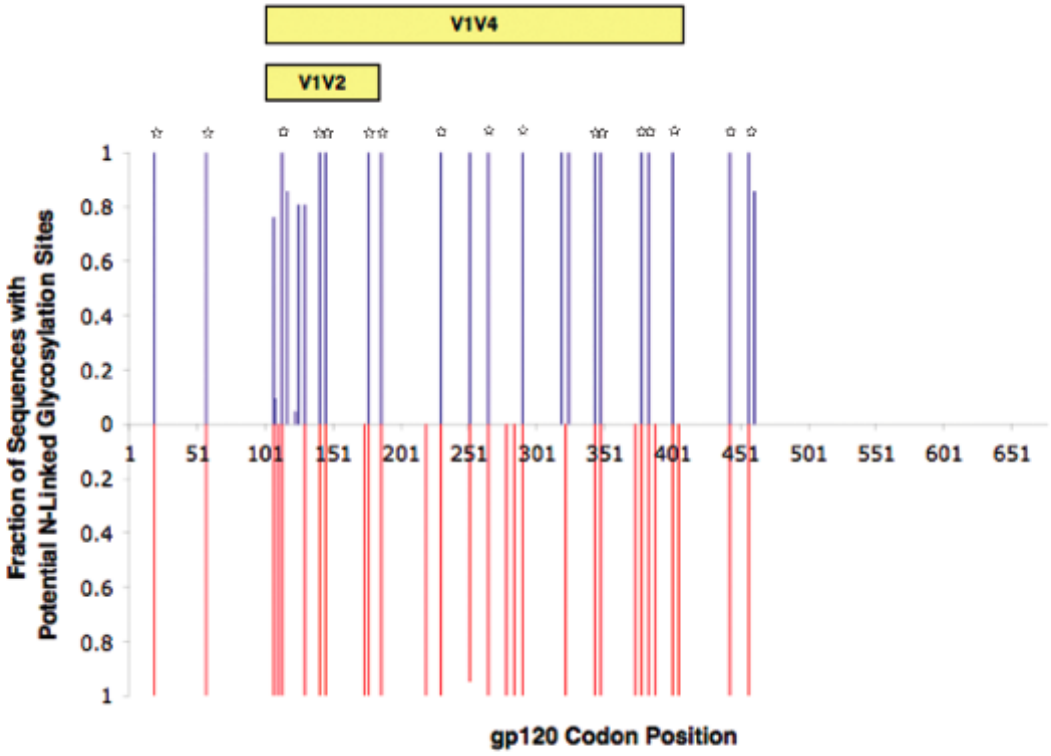


Figure 3.20 PNLGSs along SGA env gp120 AA sequences from P1 (blue) and P2 (red). *Codon position is indicated along the x-axis. The*

fraction of SGA sequences with a PNLGS at a position is indicated by the y-axis. Regions of interest are indicated by yellow boxes. Positions where 100% of sequences in both P1 and P2 have a PNLGS are indicated by small stars above the graph.

P1 had a mean of 24 PNLGSs (range 23 to 25). P2 had a mean of 29 PNLGSs (range 28 to 29). Firstly, I looked for positions where P1 and P2 were identical. P1 and P2 shared PNLGSs in 100% of sequences at 17 positions. To demonstrate that this is degree of identity was consistent with a phylogenetic cluster, I compared these sequences with 242 unrelated sequences. I studied 87 full-length, inferred amino acid sequences for gp120 sampled from other SPARTAC participants at trial baseline by population sequencing, as well as with 155 subtype B sequences from the LANL database sampled during acute infection. The combined SPARTAC/LANL reference sequences had 100% PNLGS predictions at only one site, located in C1. Greater than 90% of the reference sequences had a PNLGS at only seven positions. I concluded that the degree of similarity between P1 and P2 was consistent with a phylogenetic cluster due to transmission from a single donor.

3.4.7.2 Non-identical positions

I then looked at the positions that were not 100% identical, to see if there was any evidence of potential selection during transmission. In particular, I focussed on the V1V4 region that is implicated in susceptibility to neutralizing antibodies. Previous studies of this region have suggested that fewer PNLGSs in this region increases the susceptibility of highly related strains to neutralizing antibody (Derdeyn et al. 2004; Frost et al. 2005; Y. Liu et al. 2008; M. Zhang et al. 2004). I found a higher mean number of PNLGSs across V1V4 in P2 (24 sites, range 23-25) than P1 (19 sites, range 18-20; $p < 0.0001$,

unpaired T-test). These data suggested that there could be a difference in susceptibility to neutralizing antibodies between these two strains, consistent with a non-random model of transmission

3.4.8 Immunological analysis

3.4.8.1 No autologous or cross-neutralization observed despite potential antigenic variation

I hypothesized that differences at PNLGSs might equate to differences in neutralization that would explain the transmission of divergent *env* variants (Derdeyn et al. 2004; Frost et al. 2005; Y. Liu et al. 2008; M. Zhang et al. 2004). Therefore, I investigated whether the viral isolates from P1 and P2 had different neutralization profiles. Viruses pseudotyped with full-length day 63 *env* sequences from P1 and P2 were tested against autologous or heterologous serum from each participant sampled at day 187 post-exposure. However, the *env* pseudotypes for both P1 and P2 were only poorly neutralized or cross-neutralized (half maximal inhibitory concentration, IC_{50}, of serum ≤ 1:20, Chapter 8). Therefore it seemed unlikely that a humoral response was responsible for the detection of different *env* variants in P1 and P2, consistent with transmission being a random process.

3.4.8.2 Infectivity

However, envelope proteins are not only potentially under immune selection at transmission but also might be selected for an increased ability to enter cells. I used the data from the neutralization assay (Chapter 8) to estimate the infectivity of the env pseudotyped viruses *in vitro*. Pseudoviruses derived from P1 sequences were approximately 2.5 times ($P<0.05$) more infectious *in vitro* than pseudoviruses from P2, after normalization to reverse transcriptase

levels (Chapter 8). I noted between-host diversity in the protein sequence of C2C4, including differences in glycosylation. The C2C4 region of the gp120 protein is involved in CD4 and co-receptor binding (Rizzuto et al. 1998; Thali et al. 1991; Valenzuela et al. 1997). The *env gp120* sequences were analysed with several algorithms that were evaluated by Low and colleagues (Low et al. 2007), to detect differences in predicted co-receptor usage and minimize the possibility of missing CXCR4/CCR5 dual use variants. However, these algorithms predicted that all sampled viruses from P1 and P2 would use CCR5. The neutralization experiment was not specifically set up to test infectivity so all these results must be interpreted with caution. In addition, potential differences in infectivity do not explain why both viruses were able to cause productive infection in different individuals. Therefore, I found no evidence to reject a random model of transmission.

3.4.8.3 Class 1 HLA

I also investigated HIV-1 specific cellular immune responses, to exclude another potential source of selection that might influence my results. Clinical progression and viral load have been associated in chronic infection with host Class I HLA alleles (Brumme et al. 2008a; Carrington et al. 1999; O'Brien et al. 2001). HLA Class I restricts the ability of host CTLs to recognize and destroy infected cells. Furthermore, sequencing studies have detected evidence consistent with escape from CTL responses within weeks of HIV-1 infection (Fischer et al. 2010). The role that CTLs play in preventing established viral infection in humans remains unclear. However, vaccination of rhesus macaques to produce detectable CTL responses is associated with partial protection(Hansen et al. 2009),

and HIV-1 specific CTL responses have been detected in persons who remain PCR/ELISA negative despite high-risk exposure (Clerici et al. 1992; Langlade-Demoyen et al. 1994; Rowland-Jones et al. 1995). Therefore, I hypothesized that CTL responses during and after transmission were a potential source of selection in P1 and P2.

3.4.8.4 Between-host comparison of Class 1 HLA

Firstly, I compared the Class I HLA type of P1 and P2 with the clinical data to see if there was evidence of selection. P1 possessed HLA-A*0201, A*2402, HLA-B*1402, B*3543, Cw*0102, Cw*0802; P2 possessed HLA-A*0101, A*2901, B*0801, B*5001, Cw*0602, Cw*0701. Neither participant possessed HLA types that are strongly associated with protection from progression in chronic infection (X. Gao et al. 2001; Kiepiela et al. 2004). However, P2, who progressed quickest, possessed the HLA-A*0101 B*0801 haplotype that is associated with more rapid progression (Kaslow et al. 1990). Therefore, I hypothesized that host factors contributed to the different clinical outcome in these participants and that the viruses had been under different selection pressures since transmission.

3.4.8.5 Detected CTL responses and potential selection pressure around transmission

To further characterize these potential differences in selection due to CTL responses to gag and env peptides, I compared the clinical data to γ-interferon ELISpot data. A γ-interferon ELISpot assay measures the *in vitro* response of host CTLs to a standard panel of viral peptide epitopes (Duda et al. 2009; Frater et al. 2006; Frater et al. 2007). In chronic infection, increased numbers of responses to peptides in gag are associated with protection while an increased breadth of env responses has been associated with higher viral loads

in chronic infection (Kiepiela et al. 2007). Information on primary infection is more limited (Streeck et al. 2009). I compared significant ELISpot responses at day 63 to each participant's inferred amino acid sequence where SGA or clonal sequence data were available. At baseline P1 made responses to three OLPs in gag: ASVLSGGELDRWEKI (HXB2 p17 5-19), PIVQNLQGQMVHQAI (HXB2 p24 1-15), and VDRFYKTLRAEQASQ (HXB2 p24 165-179). By contrast, P2 responded to only one gag optimal peptide, IEIKDTKEAL (HXB2 p17 92-101), as well as a gag OLP, HQRIEVKDTLEALEK (HXB2 p17 89-103), that contained the same sequence as the gag optimal peptide. P2 also made a response to 4 env optimal peptides: TVYYGVPVWK (HXB2 gp160 30-46), SFEPIPIHY (HXB2 gp160 202-221), IVNRNRQGY (HXB2 gp160 704-712), and RRGWEVLKY (HXB2 gp160 781-795). Therefore, the ELISpot data were consistent with my hypothesis that CTL responses contributed to the different clinical outcomes in these participants and were applying different selection pressures on HIV-1 in gag.

3.4.8.6 Detectable CTL responses to not explain between-host divergence in env

I investigated whether different CTL responses could have influenced detection of divergent variants of env in my study. Phylogenetic analysis assumes neutral evolution rather than natural selection (Lemey et al. 2009). Therefore, I compared viral sequence data and ELISpot response data from each participant to see if cytotoxic T lymphocyte responses since transmission might account for observed between-host divergence in *env* (Frater et al. 2006; Streeck et al. 2009). SGA sequence data were also available for the two env gp120 optimal peptides against which P2 had a significant

response: TVYYGVPVWK (HXB2 gp160 30-46) and SFEPIPIHY (HXB2 gp160 202-221). The inferred amino acid sequences for P1 were identical to the wild-type peptides at these epitopes: TVYYGVPVWR and SFEPIPIHY. P2 was infected with both the wild-type and mutant SFEPIPIHK sequences. Clonal sequence data were available for the two env gp41 optimal peptides against which P2 had a significant response: IVNRNRQGY (HXB2 gp160 704-712), and RRGWEVLKY (HXB2 gp160 781-795). Both P1 and P2 shared the same consensus at these epitopes, which differed from the HXB2 wild-type (LVNRVRQGY and RRGWELLKY). Only a single mutant sequence away from this consensus was found in P2 at either epitope (LVNGVRQGY and RRGWELLRY). Therefore, between-host genetic differences in *env* could not be attributed to detectable, env-directed CTL responses, and my data were still consistent with transmission of *env* variants being a random process.

3.5 Discussion

3.5.1 Summary of key findings

The aim of this chapter was to study the transmissible diversity within a single host at the same point in time, or very close to the same point in time. The principle key finding in this chapter is that the genetic divergence in contemporaneously transmitted *env* variants cannot be attributed to detected immune responses within the recipient hosts. Thus the genetic divergence, which will be quantified in Chapter 4, probably represents the minimum transmissible diversity present within the unsampled donor. In reaching this finding, I have demonstrated that:

1) the sequences transmitted to P1 and P2 are highly-related consistent with the history of a common donor,

2) the *gag* sequences in particular are practically identical,

3) that differences in the *nef* and *env* sequences between P1 and P2 are not attributable to detected recipient host humoral or cytotoxic immune responses.

A second key finding is that the apparent transmission of a single *env* variant may be accompanied by the possible transmission of more than one variant in another gene, namely *pol*. In reaching this conclusion, I have demonstrated that:

1) a star-like phylogeny may be demonstrated in *env* within recently infected hosts using standard techniques, but

2) a star-like phylogeny may not be demonstrable within another HIV-1 gene (*pol*) sampled from the same participant (P2), and

3) more than one variant may be suggested by visual inspection of a viral gene (*pol*) even if the Poisson-Fitter tool suggests a star-like phylogeny, demonstrating the limitations of this tool and meaning

that a homogeneous infection should be confirmed by more than one method.

3.5.2 Sequences in P1 and P2 are highly-related

3.5.2.1 P1 and P2 sequence data consistent with recent common origin

All sequences sampled from P1 and P2 at day 63 clustered together to the exclusion of all other sequences, consistent with sharing a recent common origin. In particular, the *gag* sequences were almost identical and even sequences of the most divergent gene, *env*, clustered together. Therefore, I regarded the data as consistent with the available prior information regarding the direction of transmission. Previous studies (Hue et al. 2005; Rachinger et al. 2008; Rachinger et al. 2011; Salazar-Gonzalez et al. 2008) have used evidence of such clustering to demonstrate that sequence data are, or are not, consistent with epidemiologically-linked transmission scenarios within the same subtype.

3.5.2.2 Alternative transmission scenarios not excluded

I could not exclude the possibility that one participant infected the other because phylogenetic techniques cannot prove the direction of transmission in a forensic sense (Pillay et al. 2007). For example, I could not exclude the hypothetical possibility that two or more strains were initially transmitted to one participant, that one of these initially infectious strains was subsequently transmitted to his partner before being out-competed to extinction prior to day 63 in the initial participant.

3.5.2.3 Previous studies suggest dominant transmissible variant extinction unlikely

Extinction of the dominant transmissible variant has not been

detected in previous studies of very early infection (Bar et al. 2010; Keele et al. 2008; W. Liu et al. 2008a; J. Liu et al. 2010; Salazar-Gonzalez et al. 2008). It is not impossible that a transmissible variant may go extinct, but it would have to be replicating to a detectable level early in infection for it to be transmissible during the short window when P1 may have infected P2 and vice versa. Rather than extinction of detectable variants, previous studies have shown very early diversification of a within-host viral population of detectable, through point mutation and/or recombination (Abrahams et al. 2009; Bar et al. 2010; Goonetilleke et al. 2009; Keele et al. 2008; Keele et al. 2009a; H. Li et al. 2010; Salazar-Gonzalez et al. 2008).

3.5.3 There was no evidence that hard selection accounted for between-host divergence of variants

3.5.3.1 Overview

The chief fitness contributions of a viral gene are 1) to generate antigenic variation to evade neutralizing antibody or CTL-mediated immune responses and 2) to facilitate viral replication. Hard selection involves extinction of a variant in the host caused by the death of all viruses that belong to that variant, which would be caused by 1) the rapid appearance of a potent immune response or 2) the total inability of the virus to replicate within host cells. Thus the primary tests of a viral gene variant extinction hypothesis would involve either or both of these two mechanisms.

3.5.3.2 Previous studies addressing the hypothesis that neutralizing antibodies account for post-transmission extinction of a circulating env *variant within either host*

Previous studies had suggested that extinction or replacement of

transmitted *env* strains could be attributed to effective, though not curative, neutralizing antibody responses in both SIV (Rybarczyk et al. 2004) and HIV (Albert et al. 1990; Scarlatti et al. 1996; X. Wei et al. 2003). Previous authors had also demonstrated that transmitted env targets could be neutralized by epidemiologically linked donor serum, particularly if the number of glycosylation sites vary between donor and recipient *env* variants (Derdeyn et al. 2004; B. Li et al. 2006a; M. Li et al. 2006b). Neutralization activity was reported even in some recently-infected hosts (B. Li et al. 2006a; X. Wei et al. 2003), suggesting neutralizing antibodies may be detected in serum from the transmitting host, even if that host were himself recently infected. The appearance of neutralizing antibodies has been correlated to step-like (accelerated) accumulation of genetic distance observed in longitudinal studies of sequence diversity within recently-infected hosts (Lemey et al. 2006; Simmonds et al. 1990). This raised a further possibility, in which divergence between P1 and P2 had occurred *de novo* in either host due to immune-driven diversification in that host.

3.5.3.3 No evidence for selection due to neutralizing antibodies in P1 or P2

There was no evidence that extinction of the variants present at day 63 had been caused by a circulating neutralizing antibody response against circulating viruses in either P1 or P2. It was not possible to completely exclude the possibility that P1 infected P2 or *vice versa*, and that the transmitted variant subsequently became undetectable in the transmitting host due to a neutralizing antibody response. However, neither P1 nor P2 mounted an autologous or cross-reactive neutralizing antibody responses to pseudotyped env virions *in vitro*

using a standard assay (M. Li et al. 2005a). Furthermore, no significant neutralizing antibody was detected in P1 or P2 despite testing sera from a later time point compared with the sequence sampling time point (day 185 versus day 63). Some previous studies had reported the rapid development of neutralizing antibodies within two months of infection (Albert et al. 1990), while other studies suggest that the tempo of the neutralizing response can be subtype-specific (B. Li et al. 2006a), highly variable (Richman et al. 2003) or delayed (E. S. Gray et al. 2007a; Rong et al. 2009; Yeh et al. 2010). We expected that viruses sampled on day 63 would likely be resistant to autologous antibodies present at day 63 but that serum taken at a time point 16 weeks later (day 187) would contain detectable tires of antibody (Montefiori et al. 1996), especially if those antibodies been present at levels below detection at day 63. Previous authors had also suggested the possibility that viruses sampled close to transmission would be more susceptible to neutralization by serum taken at a time point some weeks later due to the necessity of increasing CD4 affinity at the time of transmission (Derdeyn et al. 2004; B. Li et al. 2006a; Painter et al. 2003), with CD4 affinity proposed as a potential cause for the transmission bottle-neck. However, the lack of neutralization by day 187 sera from P1 and P2, supported by the observation of between-host variation in PNGLSs and *in vitro* infectivity, suggests a more random pattern of transmission, rather than a common pattern of transmissible variation. However, it is difficult to draw firm conclusions in this regard given the small number of recipient hosts in this unique study.

3.5.3.4 Addressing the hypothesis that differences in infectivity may

account for env variant extinction

Env makes a minimal contribution to replication capacity, except in relation to the efficiency of cell entry, but the significance of this effect is difficult to quantify and the true complexity of virus-host cell interaction is not fully captured with *ex vivo* assays. Previous authors P1's env pseudotyped viruses were predicted to be less-glycosylated in gp120 and more infectious *in vitro* than P2's env pseudotyped viruses, using a particular laboratory cell-line. Viruses pseudotyped with env proteins from elite suppressors and have previously demonstrated less efficient cell entry than chronic progressors *in vitro* (Lassen et al. 2009). Replication-competent chimeras, in which *env* gene sub-domains have been substituted, also demonstrate reduced replication capacity in vitro (Lobritz et al. 2007). However, there is no evidence that cell-entry in these *ex vivo* systems replicates a meaningful selective advantage *in vivo* that would be capable of causing the rapid extinction of an *env* variant by day 63. Differences in *in vitro* viral replication capacity among replication competent strains from chronic progressors have not been consistently correlated with *env* mutation (Troyer et al. 2005), and other genes demonstrate a stronger correlation between *in vitro* differences in relative replication capacity as well as clinical outcome (Miura et al. 2009a; Miura et al. 2010). Furthermore, I reasoned that if one infers the P1 and P2 *env* variants were co-circulating to transmissible levels in the donor host, then *in vivo* differences in replication fitness were probably insignificant.

3.5.4 Differences between different viral genes in terms of between-host as well as within-host genetic divergence

3.5.4.1 Env *and* nef *were noted to have higher between-host*

divergence than mean, within-host divergence, while gag was almost identical between and within-hosts

The initial analysis of within-host divergence, presented here in Chapter 3, suggested that P1 and P2 had been infected with almost identical *gag* and *pol* sequences but quite different *nef* and *env* sequences. Published data on within-host evolution in subtype B HIV following infection of MSM over the first year of infection was very limited, represented by just two studies of *env* in the same cohort (Lemey et al. 2006; Shankarappa et al. 1999). Therefore, it was difficult to meaningfully quantify and compare the genetic distance between P1 and P2 with other studies. An additional study of longitudinally sampled env sequences had focussed specifically on only the first few days to week of infection in an analysis of 12 MSM participants but this study was of limited comparative utility because the observational time frame was shorter than in the study of P1 and P2. A more comprehensive analysis of these similarities and differences between P1 and P2 is presented separately in Chapter 4; I use coalescent, time-measured reconstructions of gene evolution to determine the significance of the apparent between-host divergence.

3.5.4.2 Evidence of multiple-variant transmission in pol within individual hosts *despite single* env *variants being transmitted to each patient*

Previous authors had suggested that neutral *env* evolution in acute infection made it the logical choice for studying viral diversification, and extrapolating single variant transmission and diversification of the whole virus (Abrahams et al. 2009; Keele et al. 2008; H. Y. Lee et al. 2009; Ochsenbauer et al. 2012). In fact, most studies in support

of the single variant transmission theory have only studied the *env* gene (Abrahams et al. 2009; Bar et al. 2010; Keele et al. 2008; M. Li et al. 2005a; H. Li et al. 2010; Ochsenbauer et al. 2012; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009; L. Q. Zhang et al. 1993).

3.5.4.3 Consideration of studies that sample multiple sequences from more than one gene at a particular time point.

However, when one considers the studies that have sampled more than one gene, the results become more mixed. One recent study that looked at two genes inferred transmission of multiple variants in *env* when only single variants had been detected in *gag*, but not *vice versa* (Novitsky et al. 2011). Another case study demonstrated a significant bottleneck at the time of transmission in env and gag (Edwards et al. 2006). However, an early study found a tendency towards greater conservation in *env* post transmission than in otherwise more conserved genes such as *gag* and *nef* (T. Zhu et al. 1993). This study would appear to be in conflict with the many studies that extrapolate whole virus evolution from neutral evolution of homogeneous *env* sequences. Yet, only a small number of patients have in fact been studied with multiple sequence sampling in more than one gene, when one pools all of the results. It therefore becomes less clear whether or not one may extrapolate whole virus evolution from multiple-sequence sampling in *env* alone. My study contributes to this limited pool of data, consistent with taking the approach that it is necessary to sample other regions of the genome before concluding that homogeneity in *env* implies single variant transmission of the entire viral genome.

3.5.4.4 Reconciling multiple pol *variants in P1 with the Poisson-Fitter result, which failed to detect multiple variants?*

Poisson-Fitter had previously been used in a number of studies to identify single-variant infections where within-host diversity had been sampled by SGA (Keele et al. 2008; H. Y. Lee et al. 2009; Novitsky et al. 2011). As previously discussed, however, these studies had also employed visual inspection as well as phylogenetic tree reconstruction to show that the sequences were consistent with a single variant infection. Where multiple-variant infections were detected by phylogenetic tree, the criteria for detecting such infections were 'extended branch lengths' for sub-populations (Novitsky et al. 2011), although what genetic distance distinguishes an 'extended' branch length from a 'short' branch length is not specified. For P1, it is likely that the majority of sequences were very close the consensus and the small sub-population representing the second variant may have undergone more or less recombination with the dominant strain, so that this small sub-population did not perturb the fitted curve sufficiently to be detected by the program. Therefore, one must question the utility of Poisson-Fitter where it is applied to ultra-deep sequencing (Fischer et al. 2010) since visual inspection or tree reconstruction is limited by the vast number of sequences.

3.6 Conclusion

3.6.1 Key Findings

3.6.1.1 Contemporaneous transmission of divergent variants consistent with a single donor rather than recipient host immune response

I have shown that highly related sequences were transmitted to P1 and P2, and relatedness was consistent with the history of a single donor. I inferred that almost identical *gag* variants were transmitted, consistent with the single donor hypothesis. Divergent *env* variants were transmitted but remained consistent with the history of exposure to a single donor. To a lesser extent, divergent *nef* variants were also transmitted but remained consistent with the single donor hypothesis. These results indicated that a single donor may be capable or transmitting significant diversity in the *env gp120* and *nef* genomic regions. In *pol*, I inferred that at least two highly related variants of *pol* may have been transmitted by the same donor to one of the recipients.

3.6.1.2 Stochastic mechanism not excluded for contemporaneous transmission of diversity

I could not account for between-host diversity by detectable immune responses using available data, so I could not refute a stochastic hypothesis for transmission of genetically divergent viral strains.

3.6.2 Concluding remarks

I demonstrated that the HIV-1 variants detected in P1 and P2 were likely to have been transmitted by a single donor to the two recipients on the same night. I go on to quantify the amount of divergence between the two recipient hosts in Chapter 4. The

immunological data indicate that currently it is not possible to predict which of the many HIV-1 variants circulating at the time of transmission will successfully seed a new infection. If transmission is a random process, than this represents a major hurdle that any HIV-1 vaccine design will need to overcome.

Chapter 4 Quantifying pre-transmission viral divergence

Parts of this chapter were published in: English, S. Katzourakis, A. Bonsall, D. Flanagan, P. Duda, A. Fidler, S. Weber, J. McClure, M. SPARTAC Trial Investigators, Phillips, R. Frater, J. (2011), 'Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants', *Retrovirology,* 8, 54.

4.1 Overview

4.1.1 Key abbreviations

Table 4.1 Abbreviations

Abbreviation	Term
95% CI/95% CL	95% confidence limit/95% confidence limit
ESS	Effective sample size
$+\Gamma$	Gamma distribution for rate heterogeneity among sites
GTR model	General time reversible model
HKY85 model	Hasegawa, Kishino and Yano 1985 model
MCMC	Markov chain Monte Carlo
SGA	Single genome amplification
tMRCA	Time of the most recent common ancestor
TN93 model	Tamura and Nei 1993 model
ULRM clock	Uncorrelated lognormal relaxed molecular clock

4.1.2 Chapter summary and hypothesis

The overarching hypothesis for this chapter is that a single donor is capable of contemporaneously transmitting significant viral diversity. In Chapter 3, I inferred that a single donor transmitted almost identical *gag* variants but that, although highly related, the *env* and nef strains were more divergent between each participant than within each participant. To demonstrate that the between-host divergence was significant, yet still consistent with the history of a

single donor, I quantified between-host diversity in each genomic region in terms of the time to the most recent common ancestor (tMRCA). I also explored some of the analytical challenges of dealing with highly related sequences and sequences that may be undergoing non-neutral evolution. I also looked at the *pol* gene, which I demonstrated in Chapter 3 may have involved the transmission of more than one variant to each participant.

4.2 Introduction

4.2.1 Quantifying transmissible diversity and implications for vaccine design

In order to be effective, a preventative vaccine design must anticipate the breadth of antigenic challenge likely to be encountered by a population of susceptible hosts (Barouch 2008; Gaschen et al. 2002; Letvin 2006; Rambaut et al. 2004). It remains unclear to what extent the circulating viral diversity within an individual human is transmissible (See Chapter 3 for a more comprehensive review of the literature on transmissibility). In Chapter 3, I presented a rare human case in which it was inferred that a single donor contemporaneously transmitted highly related but distinct viral variants to two recipients. As the donor was unsampled, it was unclear what breadth of circulating diversity in the donor was represented by the sampled sequences in P1 and P2. One way to infer the minimum transmissible diversity circulating within the donor was to quantify the estimated divergence between the variants sampled from the recipient hosts through time (Lemey et al. 2006; Lemey et al. 2009; Rachinger et al. 2011).

4.2.2 Aims

I aimed to quantify the divergence of transmissible variants sampled from each recipient in this rare case study in terms of the estimated time from transmission to the between-host tMRCA. This aim was important because if a single donor were capable of transmitting highly divergent variants, then this would have implications for the level of within-host antigenic variation that an HIV-1 vaccine design would have to anticipate.

4.3 Materials and Methods

4.3.1 Subjects

The subjects, P1 and P2, were previously described in Chapter 3.

4.3.2 Laboratory methods

Laboratory methods were described in Chapter 3.

4.3.3 Coalescent analysis methods

4.3.3.1 Tests for star-like phylogeny

See Chapter 3 for tests of star-like phylogeny.

4.3.3.2 Bayesian MCMC based analysis

I used a Bayesian MCMC based approach to quantify between host diversity in terms of the estimated time between transmission and the tMRCA. I ran BEAST v1.6.1 (Drummond and Rambaut 2007) for at least 100,000,000 steps, sampling every 10,000 steps, and employing an ULRM clock to allow for rate variation among branches (Drummond et al. 2003; Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009; Lewis et al. 2008). Convergence was assessed with Tracer v1.5 (Rambaut and Drummond 2009). Convergence was achieved if all parameter estimates had ESSs of >300 (Drummond et al. 2002; Drummond et al. 2003; Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009; Lewis et al. 2008). Viral divergence occurred significantly earlier than transmission if the confidence interval for the between-host tMRCA did not overlap with mean estimate for the day of exposure.

4.3.3.3 Demographic model

I relaxed the assumption of a constant population size by implementing the Bayesian skyline model with 10 steps for all analyses (Drummond et al. 2005; Edwards et al. 2006). I used the

model as a flexible prior on the distribution of the inter-node intervals of the sampled phylogenetic tree topologies (Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009; Lewis et al. 2008).

4.3.3.4 Tree topology

The topology of the tree was partly fixed by a within-host tMRCA prior, so that the pre-transmission branch lengths between the two hosts could be estimated. If a gene fragment demonstrated a star-like phylogeny within a recipient host, it was assumed that the tMRCA for the virus population approximated the estimated time of transmission (Drummond et al. 2003; Keele et al. 2008; Li et al. 2010). I used a normal tMRCA prior for the sequences within each participant, calibrated to a mean of 63 days since exposure (standard deviation 1 day). Although it can be noted that calibration to the time since transmission may lead to posterior overestimation of the substitution rate (Lemey et al. 2006; Lemey et al. 2009), other studies have found that this effect is small for monophyletic infections (Keele et al. 2008; Li et al. 2010).

4.3.3.5 Initial substitution model

I initially ran the tMRCA analysis with the GTR substitution model (Lanave et al. 1984a; Lanave et al. 1984b; Tavaré 1986). Convergence was not achieved with the GTR substitution model when using the within-host tMRCA estimate as the only prior; the ESSs for the prior and posterior probabilities remained <100 after 300,000,000 steps (Drummond et al. 2006; Lemey et al. 2006; Lemey et al. 2009; Lewis et al. 2008).

4.3.3.6 Normal substitution rate prior from an independent dataset

To deal with the issue of non-convergence, I estimated a prior mean rate of substitution prior for *env C2V5* by using the posterior mean

rate of an independent dataset (Drummond et al. 2003; Lemey et al. 2006). The mean rate prior was normally distributed, with a mean of 8.18 x 10^{-3} substitutions per site per year, and a standard deviation of 1.15x10^{-3} substitutions per site per year (Lemey et al. 2006). The hypervariable regions were cut to be consistent with the original dataset after consultation with the authors (Lemey 2011). To achieve convergence, my relaxed-clock analysis also required the full-length *env C2V5* fragment, rather than the part-fragment used in the reference dataset that was missing the 5´ end of the *C2* region (Lemey et al. 2006).

4.3.3.7 Robust analysis using a strict molecular clock

To determine the sensitivity of my results in *env C2V5* to the choice of parameters and priors, I also analysed the data using a strict molecular clock. For the strict clock analysis, I again used the within-host tMRCA prior to calibrate the estimated substitution rate, but I did not enforce a strong prior on the rate by using a rate estimate from the independent dataset (Lemey et al. 2006). Therefore, I could also perform this analysis for my entire 1305bp stripped *env* fragment and compare this with the results for *env C2V5*.

4.3.3.8 Robust analysis using nested substitution models

The GTR model was preferred because of it biological realism (Hue et al. 2005; Lemey et al. 2009) but it was not possible to use it for all datasets. If I did not enforce a strong prior for the substitution rate, my analysis did not achieve convergence in any of the gene fragments studied when I used the GTR model with a relaxed molecular clock. Therefore, I used two nested models within the GTR model, calibrating the time of transmission to the same prior. I used the TN93 (Tamura and Nei 1993) and HKY85 (Hasegawa et al.

1985) models. Although the TN93 model has more free parameters than the HKY85 model, both models were tested to ensure robustness of the result to the number of free parameters.

For the *nef*, *pol* and *gag* genes or fragments, both strict- and relaxed-clock analyses failed to achieve convergence after 300,000,000 steps when using the GTR substitution model. No independent reference datasets were available for calibration of evolution in these fragments. Therefore a robust analysis was carried out for both a strict and relaxed molecular clock using the nested TN93 and HKY substitution models with or without estimated rate heterogeneity among sites (+*Γ*).

4.4 Results

4.4.1 Summary of key points concerning between-host env tMRCA estimates

4.4.4.1 Between-host tMRCA estimate

By all of the tMRCA quantification analyses conducted with *env C2V5* in which convergence was achieved, the results showed that the variants that were transmitted to each participant diverged from each other significantly earlier than the time of transmission. The pre-transmission tMRCA estimates for all models used to analyse *env C2V5* are summarized in Table 4.2.

4.4.1.2 Use of the most biologically realistic model

I aimed to use the GTR model and ULRM clock together because then my analysis would include the most biologically realistic models for, respectively, nucleotide substitution and clock-like behaviour. To achieve my aim, it was necessary to include the mean rate prior from an independent dataset (Lemey et al. 2006), otherwise the analysis did not converge. The analysis also did not achieve convergence without the within-host tMRCA estimate.

Table 4.2 Pre-transmission tMRCA estimates for all models used to analyse env C2V5

Figure	Substi-tution model	Rate prior	Within-host MRCA prior	Clock	Pre-transmission tMRCA estimate, Mean (95% CIs)
**	GTR+Γ	Yes	Yes	ULRM	Did not converge
**	GTR+Γ	No	Yes	Strict	Did not converge
4.1	GTR	Yes	Yes	ULRM	2.82 years (1.28 to 4.54 years)
**	GTR	Yes	No	ULRM	Did not converge
4.2	GTR	No	Yes	Strict	2.63 years (1.14 to 4.58 years)
**	GTR	Yes	No	Strict	Did not converge
4.4	TN93	No	Yes	ULRM	2.30 years (0.180 to 4.70 years)
**	TN93	Yes	No	ULRM	Did not converge
4.5	TN93	Yes	Yes	ULRM	2.97 years (1.48 to 4.64 years)
4.6	TN93	No	Yes	Strict	2.86 years (0.983 to 5.30 years)
**	TN93	Yes	No	Strict	Did not converge
4.7	HKY85	No	Yes	ULRM	2.31 years (0.147 to 4.68 years)
**	HKY85	Yes	No	ULRM	Did not converge
4.8	HKY85	Yes	Yes	ULRM	2.97 years (1.48 to 4.63 years)
**	HKY85	Yes	No	Strict	Did not converge
4.9	TN93+Γ	No	Yes	ULRM	2.30 years (0.138 to 4.74 years)

4.4.1.3 Robust analysis

Where a mean rate prior was not used, a less parameter-rich model had to be substituted to achieve convergence for *env*. Use of a strict clock had little effect on the result, even though a strong estimated rate prior was not enforced (Table 4.2). The strong mean rate prior could be omitted with a relaxed clock if a nested substitution model was substituted for the GTR model. When the mean rate prior was not enforced with a relaxed clock, there was a much shorter estimated interval between the lower 95% CL for the tMRCA and transmission. However, the result remained significant. Thus the analysis in *env* was robust to the use of different models.

4.4.2 Quantification of pre-transmission divergence in *env*

The estimated between-host tMRCA by the most biologically plausible model (Figure 4.1, Table 4.2) was at least 1.28 years prior to transmission (mean 2.82 years, 95% CI 1.28 to 4.58 years).

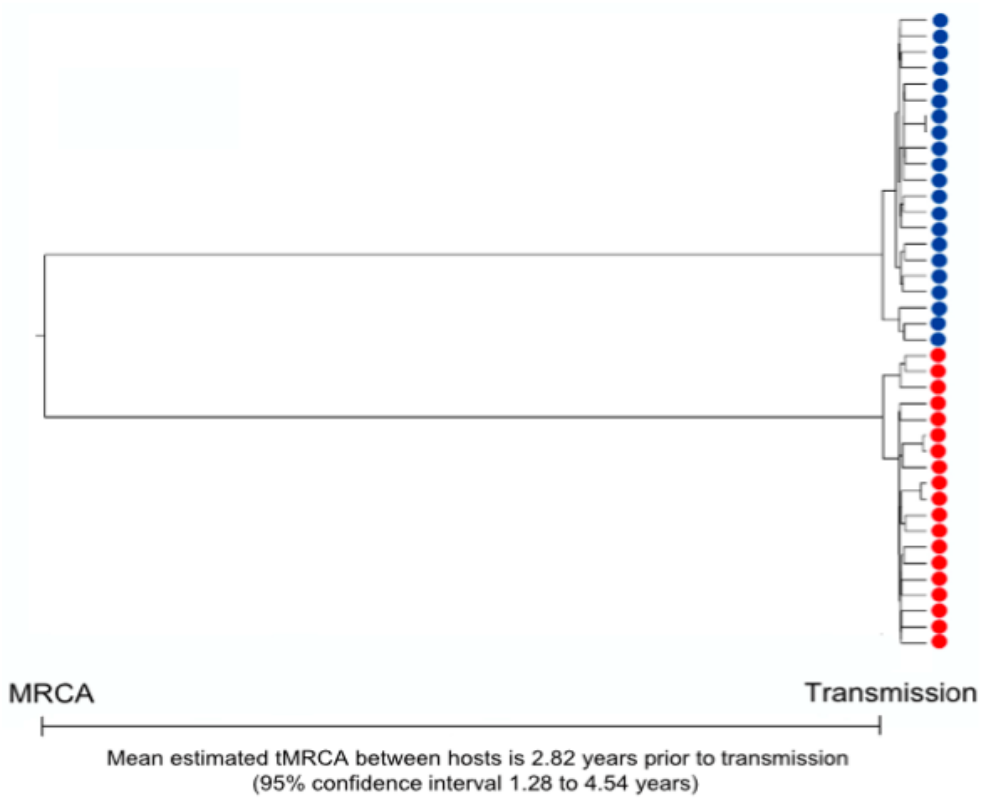


Figure 4.1 Estimation of pre-transmission divergence in* env C2V5: GTR model, ULRM clock, normal mean rate prior; blue closed dots=P1, red closed dots=P2.*

4.4.3 Robust analysis in *env C2V5* for estimating the tMRCA

Repeating the analysis without the strong mean rate prior from the independent dataset and using a strict molecular clock did not significantly alter the estimated tMRCA (Figure 4.2, Table 4.2).

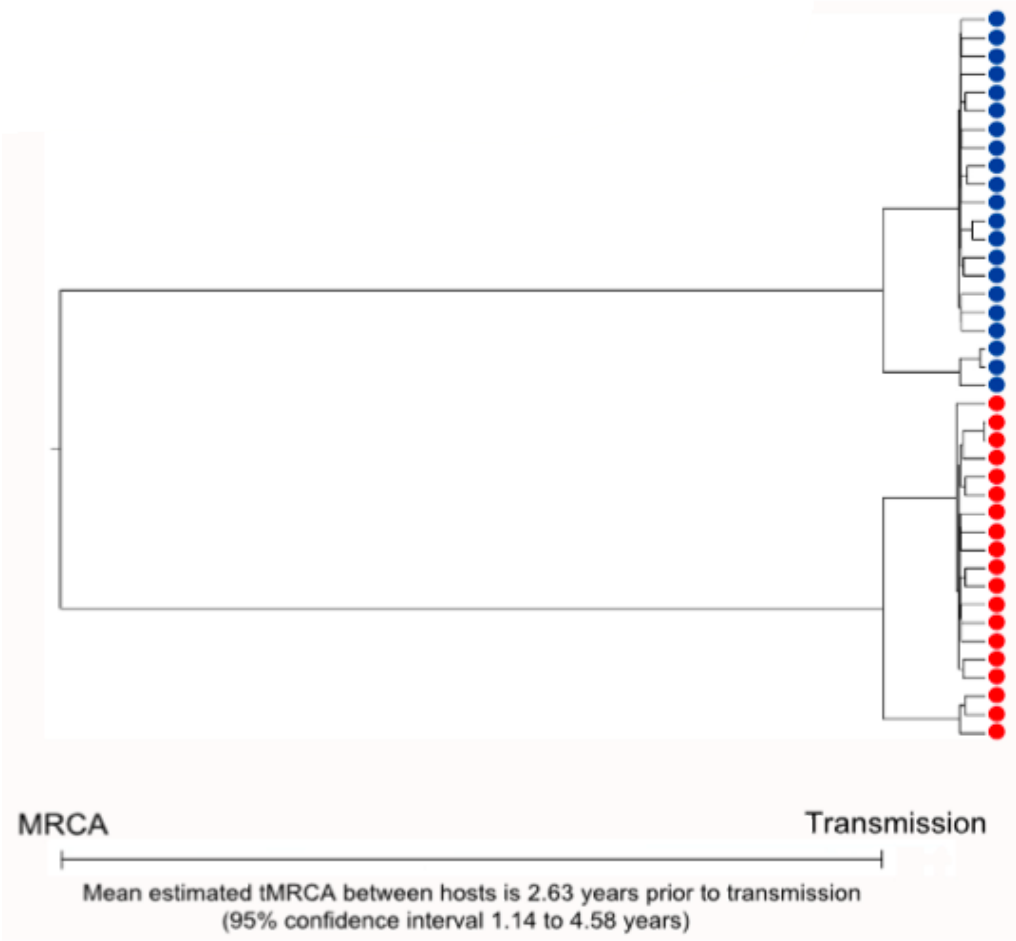


Figure 4.2* env C2V5: GTR model, strict molecular clock; blue closed dots=P1, red closed dots=P2.*

4.4.4 Robust analysis for tMRCA estimate substituting a larger *env* fragment

To investigate the sensitivity of the between-host tMRCA estimate to fragment length, I repeated the analysis using the gap-stripped alignment for the entire long *env* fragment (Chapter 3; Figure 4.3). I used a strict molecular clock, as no mean rate prior was available for the fragment and convergence was not achieved with the ULRM.

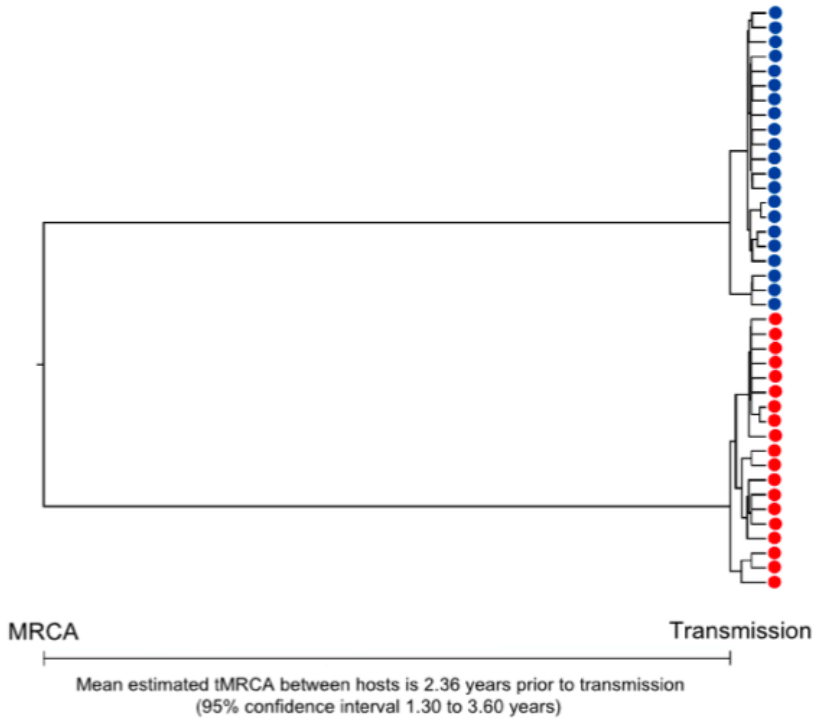


Figure 4.3 Whole* env *fragment (*gp160 *coding start to* gp120 *end): *GTR model, strict molecular clock, no mean rate prior; blue closed dots=P1, red closed dots=P2.*

The mean estimate for the time from transmission to the between-host tMRCA were not significantly altered and the lower 95% CL for the tMRCA estimate was still greater than one year prior to transmission (Figure 4.3, Table 4.2). This indicated that the results were not significantly sensitive to fragment size.

4.4.5 Quantification of pre-transmission divergence in *nef*

4.4.5.1 Convergence

In Chapter 3, it was demonstrated that the alignment for *nef* had the next highest estimate for between-host divergence after the *env* fragment alignments. As with the long *env* fragment, no substitution rate prior estimate was available from an independent dataset for *nef*. Convergence was not achieved when tMRCA estimation was attempted using the GTR substitution model with either a relaxed or strict molecular clock. In particular, the prior probability, posterior probability and G-to-T transversion parameters failed to converge.

4.4.5.2 Between-host tMRCA estimates

Analysis in *nef* proceeded using the nested TN93 substitution model, which also happens to have a single transversion parameter. The pre-transmission tMRCA estimates for all models used to analyse *nef* are summarized in Table 4.3.

***Table 4.3 Pre-transmission tMRCA estimates for all models used to analyse* nef**

Figure	Substitution model	Clock	Pre-transmission tMRCA mean estimate (95% CIs)
**	GTR+***Γ***	ULRM	Did not converge
**	GTR+***Γ***	Strict	Did not converge
**	GTR	ULRM	Did not converge
**	GTR	Strict	Did not converge
4.10	TN93	ULRM	0.833 years (0.240 to 1.57 years)
**	TN93	Strict	0.896 years (0.320 to 1.54 years)
**	TN93+Γ	ULRM	0.838 years (0.220 to 1.58 years)
**	HKY85	ULRM	0.837 years (0.225 to 1.58 years)
**	HKY85	Strict	0.876 years (0.323 to 1.50 years)
**	HKY85+Γ	ULRM	0.846 years (0.246 to 1.59 years)

As with *env C2V5*, the *nef* variants transmitted to P1 and P2 diverged significantly before transmission (Table 4.3). The tMRCA estimate in *nef* was robust to the choice of model (Table 4.3). In contrast to *env C2V5* (Table 4.2), *nef* demonstrated little variation between the 95% CI estimates between strict or ULRM clock analyses (Table 4.3). An example tree for *nef* is shown in Figure 4.10.

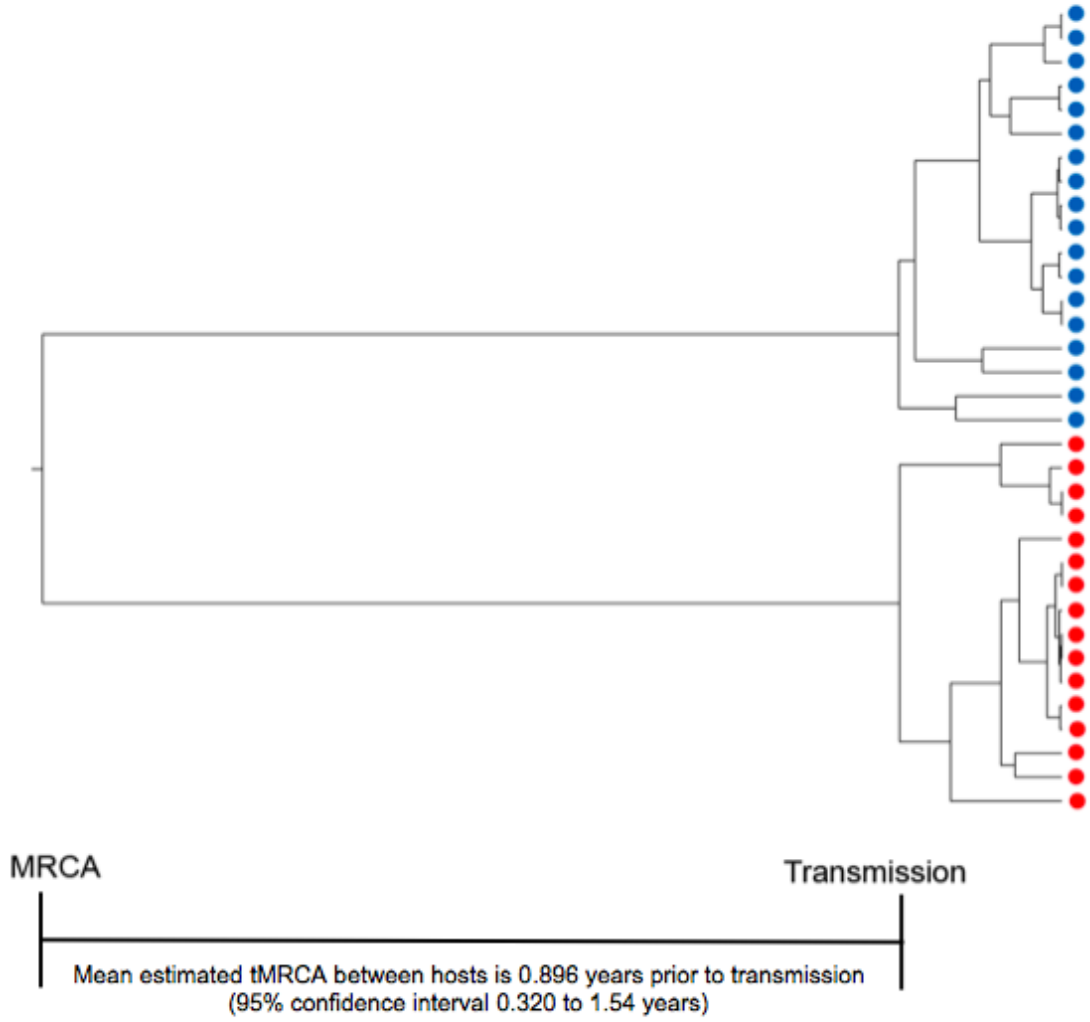


Figure 4.4 A tMRCA analysis for nef: *TN93 model, ULRM clock; blue closed dots=P1, red closed dots=P2.*

4.4.5.3 Comparing posterior substitution rates in nef *and* env

The posterior estimate for the mean rate of substitution was faster in *nef* than in *env*, when equivalent models were compared, although the difference was not significant (Table 4.4). In *nef*, the mean and upper 95% CL for the estimated mean rate using the TN93 model were less affected by using a strict clock than in *env C2V5* (Table 4.4).

***Table 4.4 Mean rate estimates for equivalent models in* nef *and* env C2V5**

Substitution model	Clock	Mean substitution rate estimate (95% CIs) x10^{-3} per site per year

TN93	ULRM	*env C2V5*: 12.0 (3.90 to 22.5) *nef*: 15.0 (7.11 to 24.2)
TN93	Strict	*env C2V5*: 10.1 (3.88 to 18.0) *nef*: 14.5 (7.43 to 23.1)
TN93+Γ	ULRM	*env C2V5*: 12.3 (3.58 to 23.8) *nef*: 15.0 (7.12 to 24.2)
HKY85	ULRM	*env C2V5*: 12.1 (4.00 to 22.6) *nef*: 15.0 (6.98 to 23.8)

4.4.6 Quantification of pre-transmission divergence in *pol*

4.4.6.1 Multi-variant infection and tMRCA analysis in pol

The main caveat to a tMRCA analysis of the *pol* fragment was the possibility that multi-variant infection had occurred in P1 (Chapter 3). Multi-variant infection would violate the assumption that the within-host tMRCA coincided with the estimated date of transmission. The expected effect would be that the analysis would produce a faster apparent mean rate of substitution in the *pol* fragment in either *env* or *nef* because more diversity would appear to have accumulated since transmission than was in fact the case.

4.4.6.2 tMRCA analysis in pol

As with *nef*, I initially the TN93 model and tested it with both strict and ULRM clock models, as well as with or without +Γ. The pre-transmission tMRCA estimates with models tested for the *pol* fragment are summarized in Table 4.5.

Table 4.5 Pre-transmission tMRCA estimates for all models used to analyse the* pol *fragment

Substitution model	Clock	Pre-transmission tMRCA mean estimate (95% CIs)
GTR+Γ	ULRM	Did not converge
GTR+Γ	Strict	Did not converge
GTR	ULRM	Did not converge

GTR	Strict	Did not converge
TN93	ULRM	0.248 years (0.000 to 0.651 years)
TN93	Strict	0.340 years (0.136 to 0.563 years)
TN93+Γ	ULRM	0.252 years (0.001 to 0.669 years)
HKY85	ULRM	0.259 years (0.000 to 0.677 years)
HKY85	Strict	0.338 years (0.136 to 0.567 years)
HKY85+Γ	ULRM	0.252 years (0.000 to 0.666 years)

The lower 95% CL for the tMRCA overlapped with the time of transmission in all relaxed clock analyses performed with the *pol* fragment (Table 4.5). Thus, according to these analyses, sequences for the *pol* fragment did not diverge significantly prior to transmission (Table 4.5). Using a strict molecular clock, the variants in P1 and P2 were estimated diverge significantly earlier than transmission, but with short intervals between the tMRCA and transmission (Table 4.5).

4.4.6.3 Posterior substitution rate estimates in pol

The substitution rate estimates for the *pol* fragment were significantly faster those for *env C2V5* (Table 4.6). The substitution rates in the *pol* fragment were significantly faster than those for *nef* when a strict clock was used. The substitution rates in the *pol* fragment were not significant or bordered on non-significance (Table 4.6). The possibility of multi-variant infection was explored by separating the two clusters of *pol* sequences observed in P1 (Chapter 3) into two taxa, giving a total of three taxa (Table 4.6) with tMRCA for each taxon calibrated to the estimated time of transmission. The analysis was repeated, using the TN93 model with either a strict or ULRM clock (Table 4.6). The results of these analyses were treated with caution because the new topology for the tree had been constrained based upon *a posteriori* information from the trees

estimated by ML and Bayesian MCMC based methods in Chapter 3. Therefore, these analyses were not used to estimate the between-host tMRCA.

Table 4.6 Mean rate estimates for equivalent models in* env C2V5, nef *and the* pol *fragment.

Substitution model	Clock	Mean substitution rate estimate (95% CIs) x10^{-3} per site per year
TN93	ULRM	*env C2V5*: 12.0 (3.90 to 22.5) *nef:* 15.0 (7.11 to 24.2) *pol*: 27.6 (14.0 to 42.6) *pol* (3 taxa): 21.1 (12.6 to 31.2)
TN93	Strict	*env C2V5*: 10.1 (3.88 to 18.0) *nef:* 14.5 (7.43 to 23.1) *pol*: 27.9 (17.3 to 39.9) *pol* (3 taxa): 20.3 (13.1 to 27.6)
TN93+Γ	ULRM	*env C2V5*: 12.3 (3.58 to 23.8) *nef:* 15.0 (7.12 to 24.2) *pol*: 27.7 (15.2 to 44.2)
HKY85	ULRM	*env C2V5*: 12.1 (4.00 to 22.6) *nef:* 15.0 (6.98 to 23.8) *pol*: 28.3 (15.3 to 45.2)

The analysis was repeated, using the TN93 model with either a strict or ULRM clock (Table 4.6). The results of these analyses were treated with caution because the new topology for the tree had been constrained based upon *a posteriori* information from the trees estimated by ML and Bayesian MCMC based methods in Chapter 3. Therefore, these analyses were not used to estimate the between-host tMRCA. The posterior estimate for the substitution rate in the *pol* fragment assuming three taxa was no longer significantly quicker than for the equivalent model in *env* or *nef* if using a ULRM clock (Table 4.6). For the strict clock analysis, the estimated substitution rate remained significantly higher in the *pol* fragment than *env C2V5* but not *nef* (Table 4.6).

4.4.7 Quantification of pre-transmission divergence in *gag*

4.4.7.1 tMRCA analysis in gag

The pre-transmission tMRCA estimates for the *gag* fragment using different models are summarized in Table 4.7.

Table 4.7 Pre-transmission tMRCA estimates for all models used to analyse the* gag *fragment

Substitution model	Clock	Pre-transmission tMRCA estimate Mean (95% CIs)
GTR+Γ	ULRM	Did not converge
GTR+Γ	Strict	Did not converge
GTR	ULRM	Did not converge
GTR	Strict	Did not converge
TN93	ULRM	0.000 years (-0.005 to 0.024 years)
TN93	Strict	0.007 years (-0.004 to 0.028 years)
TN93+Γ	ULRM	0.007 years (-0.004 to 0.026 years)
HKY85	ULRM	0.006 years (-0.004 to 0.026 years)
HKY85	Strict	0.006 years (-0.003 to 0.026 years)
HKY85+Γ	ULRM	0.006 years (-0.004 to 0.025 years)

The tMRCA between hosts was estimated to be almost identical to the tMRCA prior estimate for within-host evolution (Table 4.7). Therefore, it was inferred that P1 and P2 were infected with almost identical *gag* variants at the time of sampling.

4.4.7.2 Posterior substitution rate estimate

The estimated substitution rate for *gag* was very similar regardless of the model used (Table 4.8). The substitution rate estimates for *gag* were also very similar to those for *env C2V5*, when equivalent models were used (Table 4.8). It was also noted that *gag* estimates had narrower 95% CIs than the *env C2V5* estimates.

Table 4.8 Mean rate estimates for equivalent models in the* gag *fragment and* env C2V5*.

Substitution model	Clock	Mean substitution rate estimate (95% CIs) x10^{-3} per site per year
TN93	ULRM	*env C2V5*: 12.0 (3.90 to 22.5) *gag*: 12.5 (9.12 to 16.6)

TN93	Strict	*env C2V5*: 10.1 (3.88 to 18.0) *gag*: 12.5 (8.71 to 16.1)
TN93+Γ	ULRM	*env C2V5*: 12.3 (3.58 to 23.8) *gag*: 12.6 (9.04 to 16.4)
HKY85	ULRM	*env C2V5*: 12.1 (4.00 to 22.6) *gag*: 12.5 (8.72 to 16.2)

4.5 Discussion

4.5.1 P1 and P2 were infected with very different *env* variants.

The estimated between-host tMRCA for the *env* fragment, using all available prior information, was greater than one year prior to transmission. The only models in which the estimated tMRCA was less than one year prior to transmission were the ULRM clock models in which no mean rate prior was specified. Thus for ULRM clock analyses, the lower 95% CL for the estimated substitution rate tended to be faster without the mean rate prior than with it. However, the result was still statistically robust to omission of the normal rate prior.

4.5.2 E*nv* fragments contained enough phylogenetic information for between-host tMRCA estimation using the GTR model

It was noted that *env C2V5* and the full *env* (*gp160* coding start to *gp120*) fragment were the only regions of the HIV-1 genome in which convergence was achieved with the most parameter-rich substitution model, the GTR model, without specifying a substitution rate prior. Assuming monophyletic infection, these analyses were consistent with relatively neutral evolution by nucleotide substitution in the context of high viral replication during primary infection (Keele et al. 2008).

4.5.3 Transmission of significant *nef* diversity

4.5.3.1 tMRCA estimation consistent with different nef *variants*

The estimated between-host tMRCA for *nef* was significantly earlier than the estimated date of transmission. However, the estimated tMRCA was not as early for *nef* as for *env C2V5*. No mean rate prior was available for *nef*, so the estimated substitution rate could not be

calibrated with an independent dataset.

4.5.3.2 Competing selection pressures in nef *could account for more recent estimated tMRCA between hosts*

Competing selection pressure to the more recent tMRCA estimate in *nef*. Functional conservation would reduce the level of detected divergence (Holmes 2009; Lemey et al. 2009). Diversifying selection due to early cellular immunity in P1 and P2 could lead to overestimation of the rate of substitution (Brumme et al. 2008b). In addition, the *3´LTR* overlaps its transcription factor binding region with roughly 50% of *nef* (Deacon et al. 1995). The transcription factor binding region is important for viral replication rate, so one would expected it to be functionally conserved (Centlivre et al. 2005). As the donor remained unsampled, it remains unknown whether *nef* was under similar competing selection pressures prior to transmission. Despite these competing selection pressures, the results remained consistent with divergent but transmissible *nef* variants circulating within a single donor.

4.5.4 P1 and P2 were infected with identical *gag* variants

4.5.4.1 Gag divergence consistent with a single donor

The estimated between-host tMRCAs for the *gag* fragment did not occur significantly before the time of transmission. This was consistent with a single donor as the source for the infections in P1 and P2.

4.5.4.2 Presence of identical gag variants suggestive of inter-fragment recombination

Very different *env* variants were associated with almost identical *gag* variants, consistent with selective constraint, recombination or both in the previous donor (Lamers et al. 2009). Recombination is already known to occur at a high rate among HIV-1 variants *in vivo* (W. S. Hu and Temin 1990a, 1990b). Recombination and selection would be consistent with HIV-1 genes, or genetic regions, behaving as independent agents that transiently co-locate within a virus (Gardner and Grafen 2009).

4.5.5 Evidence consistent with multi-variant *pol* transmission

4.5.5.1 Posterior substitution rate estimation

The significantly faster substitution rate that was estimated in *pol* compared with that in *env C2V5* could be consistent with multiple variants being transmitted by the donor to P1. The estimated to rate may also be increased if multiple transmitted variants were transmitted (Chapter 3) and underwent recombination (Lemey et al. 2004). Since *pol* sequences were sampled by bacterial cloning of population-amplification sequences, it cannot be excluded that recombination occurred *in vitro* (Jordan et al. 2010; Salazar-Gonzalez et al. 2008). An alternative explanation for a rapid substitution rate would be rapid diversifying selection pressure post-transmission (S. L. Liu et al. 1997; D. J. Wilson and McVean 2006).

4.5.5.2 Attempts to correct for multi-variant transmission

Separating the two variants evident in P1 (Chapter 3) did lower the substitution rate. However, caution is needed in interpreting this result as the taxa were defined *a posteriori* from the topology of

unrooted trees estimated by ML and Bayesian MCMC based methods. A better approach would be to apply a strong rate prior without having to calibrate the substitution rate to the estimated within-host tMRCA (Lemey et al. 2009). For example, one might apply a prior on the substitution rate or serially sample the donor, but neither of these options was available for this analysis.

4.6 Conclusion

4.6.1 Key Findings

In this chapter, it has been demonstrated that quantification of the tMRCA in epidemiologically-linked but genetically divergent *env* variants was consistent with evolution over at least a year in a chronically infected donor. Furthermore, the between-host tMRCA for *nef* also diverged significantly earlier than transmission by the models tested. The recipients were infected with virtually identical *gag* variants, consistent with a single donor. Lastly, multi-variant pol transmission with or without recombination may have occurred in P1. Therefore, a single donor may be capable of contemporaneously transmitting viruses with significant genetic diversity.

4.6.2 Suggestions for further work

4.6.2.1 Replicating this case study

To validate the results of this chapter, it would be necessary to identify more cases of contemporaneous transmission of significant diversity by single donor to multiple recipients. However, the case study presented in Chapter 3 and 4 is particularly rare because of the combination of clinical and sequencing data. In Chapter 3, another pair of recipients with epidemiologically linked infections were presented but inferring a single donor *a priori* from the clinical data was more problematic. Therefore, replicating the data presented in Chapters 3 and 4 would be challenging.

4.6.2.2 Obtaining prior rate estimates

Obtaining prior rate estimates for *nef*, *pol* and *gag* from independent datasets would allow calibration of the within-host substitution rates for these genes or fragments. Independent

datasets would also provide some degree of validation for the substitution rate estimates obtained with the analyses, particularly where multi-variant transmission was suspected, such as in *pol*.

Chapter 5 Inferring intra-subtype HIV-1 multiple-infection

5.1 Overview

5.1.1 Brief summary of background to chapter

Designs for a preventative HIV-1 vaccine would be enhanced by understanding the mechanism of transmission and strategies that the virus has developed to evade eradication by the immune response. Previous authors have suggested that, in the vast majority of HIV-1 cases involving sexual transmission, all viral diversity within the host evolves from a single founder strain, or even virus (Abrahams et al. 2009; Bar et al. 2010; Keele et al. 2008; Keele et al. 2009a; Lee et al. 2009; Li et al. 2010; Ochsenbauer et al. 2012; Salazar-Gonzalez et al. 2008; Salazar-Gonzalez et al. 2009). However, viral growth kinetics and evolution may mask initial viral diversity during and after the exponential growth phase. Sampling the exact population of transmitted virus is currently impossible in human beings for technical and ethical reasons (Haase 2005; Keele et al. 2009a; Ochsenbauer et al. 2012; Pope and Haase 2003).

5.1.2 Brief summary of chapter and hypothesis

Therefore, the overall aim of Chapters 5 and 6 is to use epidemiologically unlinked multiple-infection in early chronic infection as a surrogate to demonstrate that persistent co-circulation of, and recombination between, divergent HIV-1 variants may have important implications for viral adaptation to human hosts. In Chapter 5, I hypothesized that P1 had been infected by an epidemiologically unlinked strain and that I evaluate the

circumstantial and laboratory evidence for multiple-infection in participant P1. I demonstrated that P1 becomes infected with an epidemiologically unlinked *nef* strain by day 235 post initial infection. However, complete replacement by an epidemiologically unlinked strain was not seen in *gag* or *env*. I evaluated the immunological data for the first 474 days of P1's infection to see if the strains present at later time-points possessed a fitness advantage in terms of immune-escape. I concluded that the variants present at later time points did not possess a clear immunological advantage. I went on to explore why such a pattern of viral variant replacement may arise in the absence of an immunological explanation in Chapter 6.

5.2 Introduction

5.2.1 Multiple-infection

5.2.1.1 Detecting multiple-infection

HIV-1 does not appear to produce a natural immune response that absolutely prevents multiple-infection with epidemiologically unlinked strains, even within the same viral subtype (Altfeld et al. 2002; D. M. Smith et al. 2005a) yet published reports of intra-subtype multiple-infection are relatively rare even in high-risk groups (Rachinger et al. 2011b; D. M. Smith et al. 2005a; van der Kuyl and Cornelissen 2007). Two possible explanations for the rarity of multiple-infection are that viruses have different relative fitnesses and that the invading strain must grow from a tiny initial relative frequency within the viral population (Fisher 1922, 1930). A further possibility is that within-subtype infections are genetically very similar, making screening for multiple-infection difficult. In fact, many cases studies of multiple-infection have only been detected because researchers investigated in depth why a host experienced an unexpected acceleration in disease progression or a spike in viral load (Altfeld et al. 2002; Gottlieb et al. 2004; Rachinger et al. 2008; van der Kuyl and Cornelissen 2007).

5.2.1.2 Relevance of multiple-infection to primary infection

Studies of epidemiologically unlinked multiple-infection have implications for studies of acute primary infection. During acute infection, a single dominating strain is capable of producing a viral load of over 10^7 per ml (Dalmau et al. 2012; Keele et al. 2008) and, due to its high density, may frequently recombine with other variants within the host, particularly in lymph nodes (Jung et al. 2002). If initial viral diversity is masked by the exponential growth

phase and early dominance by a single strain, the consensus sequence for viruses detected during acute infection may not fully reflect the true breadth of transmitted diversity (Derdeyn et al. 2004).

5.2.2 Transmission, multiple-infection and the immune response

It is of particular interest to researchers of an HIV-1 vaccine to know whether or not both the relative infrequency of co-infection and multiple-infection in HIV-1 is due to a modifiable host immune system barrier (Haase 2005; Keele et al. 2008; Ochsenbauer et al. 2012; Pope and Haase 2003). The main qualitative difference between studying multiple-infection in the chronic phase and co-infection during the acute phase is the relative development of the host immune response. The immune response is likely to have an impact upon relative viral fitness, such that there is an adaptive trade-off between immune evasion and replication rate in either strain (Brockman et al. 2010; Brumme et al. 2011; Miura et al. 2009a). During the acute phase, replication rate may be the most important determinant of relative fitness whereas in chronic infection, an ability to avoid an immune response might be more important for relative viral fitness (Altfeld et al. 2002; Streeck et al. 2008).

5.2.3 Detection versus inference of multiple-infection within a subtype

Detection of within-subtype multiple-infection, even with epidemiologically unlinked strains, remains problematic due to the relatively high degree of sequence similarity (Truszkowski and Brown 2011; van der Kuyl and Cornelissen 2007). However, if the degree of similarity between sequences is higher then the expected

rate of homologous recombination may also be higher, due to restrictions on the mechanisms of recombination (Chin et al. 2005; Nikolaitchik et al. 2011), and recombination following multiple-infection may be of great evolutionary importance (Streeck et al. 2008). However, multiple infection is usually inferred indirectly by the detection of a super-infecting strain, often because detection of an epidemiologically unlinked strain is attempted following unexpected clinical deterioration in a chronically infected host (Altfeld et al. 2002; Koelsch et al. 2003; Rachinger et al. 2008; Rachinger et al. 2011a; D. M. Smith et al. 2005b; Streeck et al. 2008; Truszkowski and Brown 2011). Furthermore, the genes *env gp120*, *gag p17* or *gag p24* are frequently sampled to detect superinfection (Rachinger et al. 2008; Rachinger et al. 2011a; Streeck et al. 2008) but superinfection may be missed if mechanisms of homologous recombination are restricted by mutation in these particular genes (Chin et al. 2005; Nikolaitchik et al. 2011). In addition, one cannot discount primer-specific amplification bias when attempting to amplify divergent strains in multiple-infection of the same subtype using any of the currently standard techniques (Jordan et al. 2010).

5.2.4 Points addressed in this chapter

In this Chapter, I used population sequencing from multiple time points over the first 819 days from primary infection to infer epidemiologically unlinked, intra-subtype multiple-infection before day 235 in P1, who was one of the recipient hosts presented in Chapters 3 and 4. In addition I used available immunological data from the first 474 days of infection to explore the evolutionary significance of multiple-infection in P1. If multiple-infection could persist for several months within P1 and ultimately lead to selectively advantageous superinfection, then it was possible that epidemiologically linked co-infecting strains could also persist through the first few days to weeks of the initial infection and contribute significantly to the evolution and adaptation of the virus to the host's immune response.

5.3 Materials and Methods

5.3.1 Subjects

The subjects, P1 and P2, were previously described in Chapters 3 and 4 of this thesis. P1 and P2 were a couple, and presented to trial follow-up appointments on the same day. P1 was frequently sampled before and after the inferred appearance of the epidemiologically unlinked strain because of his enrolment in the SPARTAC trial, rather than because of his clinical or multiple-infection status. Eight pre-treatment time-points were available for sequencing in P1. These time points, given as days post exposure to the original common donor, are shown in Table 5.1, with the SPARTAC trial protocol visit title given as weeks post-enrolment. Consecutive pre-treatment time points were also available from P2 for days 63, 137, 185 and 235.

Table 5.1 Time points sampled from P1

Days post-exposure	SPARTAC trial protocol visit (week post-enrolment)
63	0
137	12
185	16
235	24
397	48
439	52
474	60
819	108

5.3.2 Laboratory methods

5.3.2.1 Blood collection and plasma separation

Blood samples were collected from the two participants at each pre-treatment time-point as described in Chapter 3. Plasma was

separated and stored as described in Chapter 3.

5.3.2.2 Generation of cDNA

Viral RNA was extracted from defrosted plasma and used to generate cDNA by reverse transcription, as described in Chapter 3. Extracted RNA and cDNA were stored as described in Chapter 3.

5.3.2.3 Population-amplification and sequencing

Population-amplification and sequencing were undertaken for *gag*, *env* and *nef*. Protocols for population-amplification and sequencing from undiluted cDNA template were described in Chapter 3. The analysed *env* fragment was the same as the long *env* fragment in Chapter 3. The analysed *nef* fragment was the same as the *nef*-only fragment in Chapter 3. The *gag* fragment was shorter than the long gag fragment in Chapters 3 and 4 because multiple sets of primers were used for PCR. Primers are listed in the Appendix (Chapter 8).

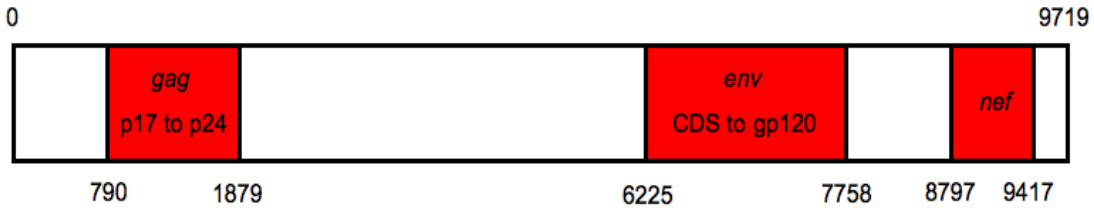


Figure 5.1 Regions of the HIV-1 genome analysed in Chapters 5 and 6

Amplification and sequencing of *gag* was attempted for all pre-treatment time points in each participant. Amplification and sequencing for *env* and *nef* were attempted at all time points except post-exposure day 185.

5.3.2.4 Steps taken to limit PCR-induced errors during population-amplification

The precise identity of the individual sequences was less of a

priority than detecting multiple variants within a sample but I took steps to limit the effects of PCR-induced recombination on my analysis. For *gag*, I repeated the PCR amplification with and without different primers and at multiple time points to try and limit the chance effects of *in vitro* errors, such as misincorporation, recombination and primer bias (Clarke et al. 2001; Erlich et al. 1991; Jordan et al. 2010). Furthermore, I used PCR extension times at the upper limit of the manufacturer's recommendations for the size of the target template (Erlich et al. 1991). For *env*, involving the longest first-round product (4.5 to 5 kb), I used a high-fidelity polymerase enzyme (Invitrogen, UK) that can amplify fragments up to 15kb as well as the manufacturer-recommended extension time of 5 minutes (1 minute/kb). However, as I was analysing population PCR sequences, I would only be obtaining a consensus, and so PCR methods would not avoid this issue.

5.3.3 Phylogenetic Analysis

5.3.3.1 Sequence alignment

Gag, *env* and *nef* sequences were edited, aligned and gap-stripped as described in Chapter 3. Sequences were aligned to population-amplification sequences obtained from SPARTAC participants as described in Chapter 3. Where appropriate, additional reference sequences were obtained from the LANL HIV sequence database.

5.3.3.2 Phylogenetic tree reconstruction by ML estimation

Phylogenetic trees were initially constructed using the maximum likelihood (ML) method with PhyML v3.0 software (Guindon and Gascuel 2003) and visualized in FigTree v1.3.1 (Rambaut 2009). The chosen substitution model gave the highest likelihood when screened with jModelTest (Posada 2008, 2009). The chosen model

was the generalized time reversible (GTR) model (Lanave et al. 1984a; Lanave et al. 1984b; Tavaré 1986) incorporating estimates of the proportion of invariant sites (I) and the shape parameter (α) of a gamma (Γ) distribution (Büchen-Osmond). Phylogenetic distance estimates were obtained using Mega5 (Tamura et al. 2011) and TREE-PUZZLE v5.3.rc7! (Schmidt et al. 2002; Schmidt and von Haeseler 2007).

5.3.4 Recombination Analysis

5.3.4.1 Visual assessment

I visually screened unstripped nucleotide sequences for the fragments from all three genes using Highlighter v2.1.1 (LANL, US), with the day 63 (baseline) consensus for P1 used as the master sequence for comparison.

5.3.4.2 Distance-based screening for recombination breakpoints

A query-versus-reference approach was taken for distance-based inference of recombination. The sequences were screened for break-points using RIP v3.0 (LANL, US), by comparing sequences against the day 63 (baseline) consensus for P1 as well as the most divergent strain in P1. The window size used was 100 base pairs, which was the smallest allowable window size. The statistical-significance threshold for detection for detection of recombination was a p-value of less than 0.05. Recombination was confirmed using the method implemented in SimPlot v3.5.1 (Lole et al. 1999). In addition to pairwise distance, this method infers recombination by estimating phylogenetic clustering along the alignment, with statistical support reported in terms of the percentage of permutated trees after 1000 bootstrap replicates. In SimPlot, the window size used was 60bp and the phylogenetic model used for the bootstrapping analysis was the

Kimura 2-parameter + neighbour joining (NJ) model.

5.3.4.3 Alternative algorithm-based confirmation of recombination breakpoints

3Seq analyses all recombination mosaic possibilities of triplet sequences (Boni et al. 2007; Hogan and Siegmund 1986; Lemey et al. 2009) rather than using a query-versus-reference approach. The window size used was 40bp.

5.4 Results

5.4.1 Clinical evidence for exposure of P1 to a fourth host after day 63

P1 and P2 were sampled on the same day when they enrolled in the SPARTAC trial, 63 days post-exposure to the original, common donor. Both participants were randomized to receive no therapy. P1 reported exposure to a fourth party after day 63 and before day 235. Evidence of HIV-1 super-infection in P1 was seen on plasma collected at day 235. On day 245, P1 was diagnosed with acute hepatitis C virus (HCV) infection, having been negative for HCV by PCR and antibody on day 29. He commenced treatment with ribavirin and interferon after day 245. P2 remained HCV negative.

5.4.2 Sequence analysis

5.4.2.1 Summary of results of sequence analysis

Multiple-infection of P1 between days 63 and 235 was initially inferred from the separate clustering of *nef* by ML estimation, with bootstrap support of greater than 99% (Figure 5.2). By contrast, the P1 and P2 clusters for *gag* and *env* continued to cluster together to the exclusion of all other sequences, with bootstrap support of 97.6% and 100% respectively (Figures 5.3 and 5.4). However, the degree of within-host divergence for P1 in *gag* from day 235 appeared to be larger than expected. Furthermore, there was evidence of intra-subtype recombination at days 235 and 439 in *gag*, day 474 in *env*, and day 819 in *nef*.

5.4.2.2 Maximum likelihood estimation for nef

In *nef*, population-amplification from P1 and P2 sampled on days 63 and 137 formed one cluster (N1) while P1 sequences from days 235, 474 and 819 formed a separate cluster (N2), with bootstrap support of greater than 99% (Figure 5.2). Both clusters sat within the larger subtype B cluster. The two clusters were separated by other sequences from the UK cohorts, indicating that P1 had been multiply infected with sequences from epidemiologically unlinked donors and that subsequent replacement of the major variant (superinfection) had taken place.

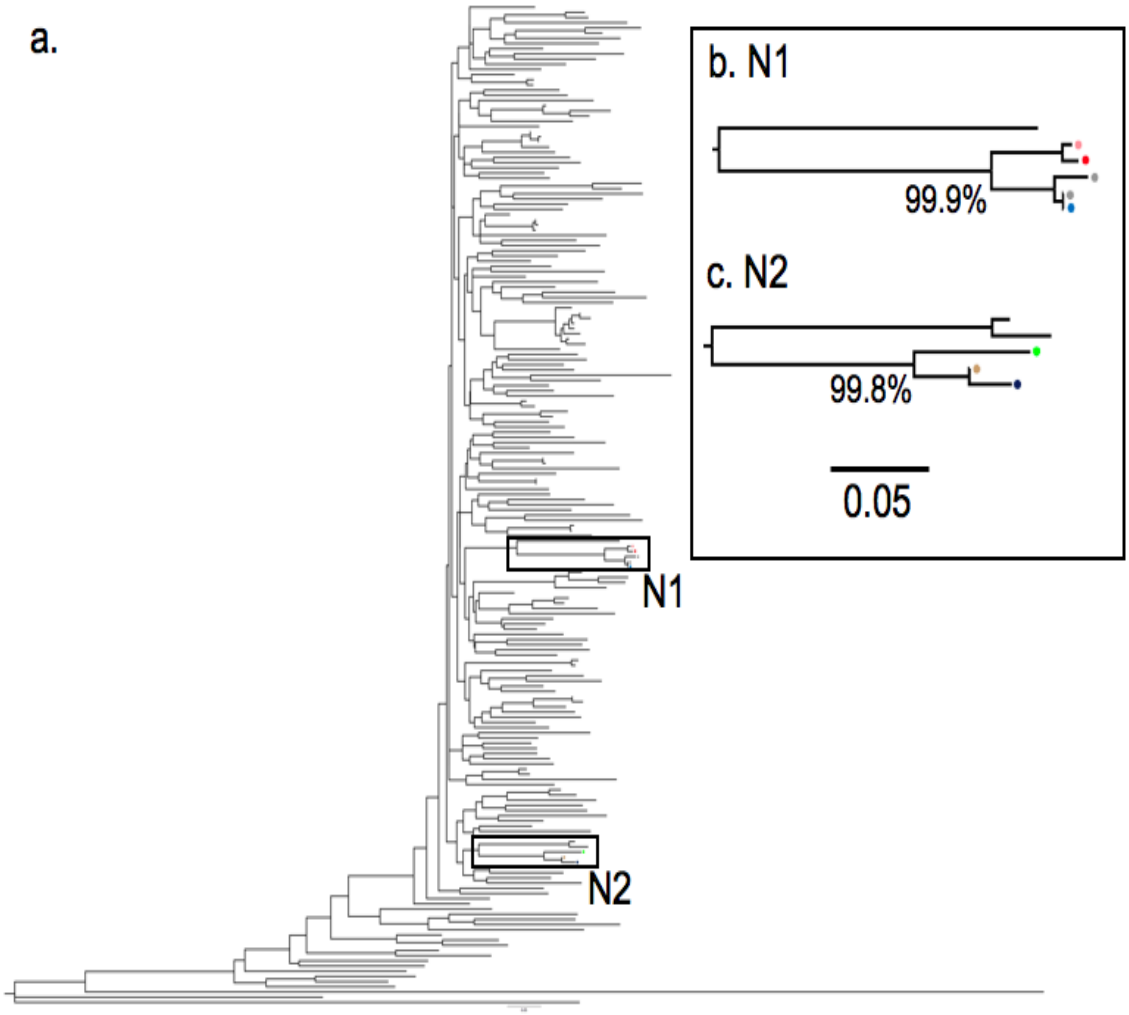


Figure 5.2 The a. full ML tree for* nef *population amplification sequences and zoomed in views of the b. original sequences

cluster (N1) and c. the new P1 sequences cluster (N2). *For P1, the colours are: intermediate blue=day 63, grey=day 137, dark blue=day 235, brown=day 474, green=day 819. For P2, the colours are: light red=day 63, pink=day 137, dark red=day 235. Bootstrap support for clusters in shown as a percentage. The scale bar represents genetic distance, so that 0.05=5%. The nearest branches with the nearest sequences that are neither from P1 nor P2 are shown for comparison.*

5.4.2.3 Maximum likelihood estimation for env

All *env* population-amplification sequences from P1 and P2 clustered together to the exclusion of all other sequences with bootstrap support of 100% (Figure 5.3). Furthermore, all *env* for P1 clustered together with bootstrap support of 99.6%. Therefore, *env* data were not suggestive of multiple-infection.

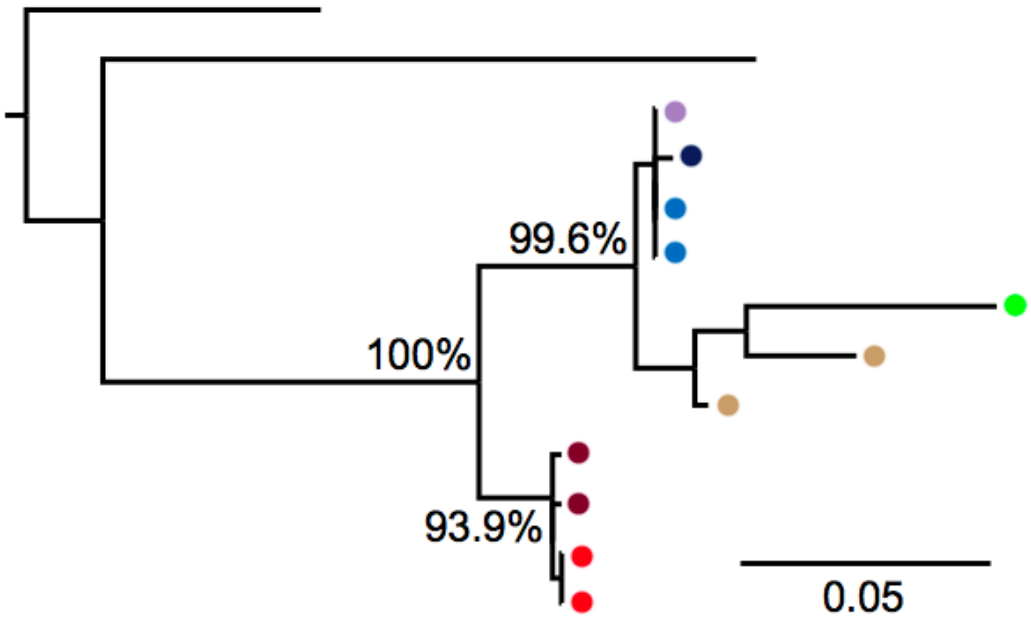


Figure 5.3 ML tree for* env *population amplification sequences. *For P1, the colours are: intermediate blue=day 63, dark blue=day 235, purple=day 439, brown=474 and green=day 819. For P2, the colours are: light red=day 63 and dark red=day 235. Bootstrap support for clusters in shown as a percentage. The scale bar represents genetic distance, so that 0.05=5%.*

5.4.2.4 Maximum likelihood estimation for gag

Analyses of the population-amplification sequences from *gag* were

consistent with multiple-infection in P1 from day 235 (Figure 5.4) with very similar sequences. All sequences in P1 and P2 clustered together with bootstrap support of 97.6%.

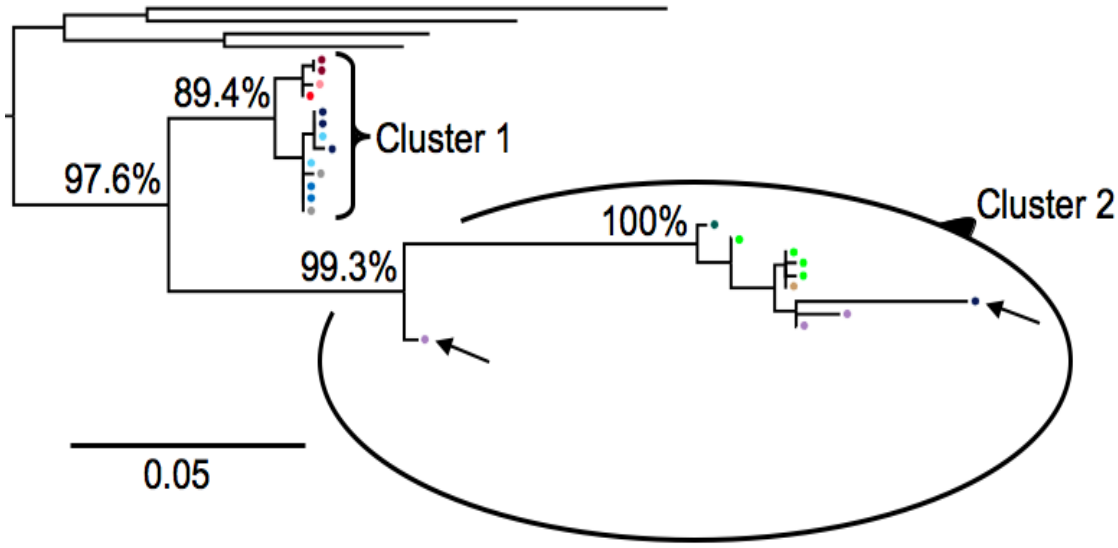


Figure 5.4 ML tree for* gag *population amplification sequences. *Bootstrap support for clusters are shown as percentages. The scale bar represents genetic distance, so that 0.05=5%. Cluster 1 and Cluster 2 are indicated. Arrows point top the suspected recombinant sequences. For P1, the colours are: intermediate blue=day 63, grey=day 137, light blue=day 185, dark blue=day 235, dark green= day 397, purple=day 439, brown=day 474, green=day 819. For P2, the colours are: light red=day 63, pink=day 137, dark red=day 235.*

Population-amplification sequences for *gag* from P1 sampled on days 63, 137, 185 and 235 formed a cluster (Cluster 1) with sequences from days 63 and 137 in P2. Population-amplification sequences for *gag* from P1 sampled on days 235, 397, 439, 474 and 819 formed a second cluster (Cluster 2) with a bootstrap of 99.3%. Cluster 2 was not separated from the first cluster by other sequences within the database. Sequences from day 235 were found in both clusters. A sequence from P1 at day 439 sat at the base of the second cluster, some distance from the other sequences from P1

that formed the nearest main cluster.

5.4.2.5 Sequence diversification in gag

Two potential recombinant sequences were removed and the mean pairwise distance was calculated by composite ML estimation in Mega5 (Tamura et al. 2011). The mean distance within Cluster 1 was 0.1% (standard error 0.1%) and the mean distance in Cluster 2 was 1.3% (standard error 0.3%). By contrast, the mean distance between Clusters 1 and 2 was 5.0% (standard error 1.2%). Percentage divergence was calculated by ML estimation in TREE-PUZZLE v5.3.rc7! (Schmidt et al. 2002; Schmidt and von Haeseler 2007). The greatest divergence in the combined P1/P2 ML analysis of *gag* was found in P1, particularly between the sequences identical to those sampled at day 63 and a sequence sampled from day 439 (6.238%). The maximum divergence between all sequences sampled in P2 up to day 235 was 0.19%, whereas for P1 it was 4.4%.

5.4.2.6 Comparison with a previous study of gag *evolution and diversification in chronic infection*

A previous study was consistent with the amount of diversification observed in P2 but not P1. A previous study (Piantadosi et al. 2009) estimated that within-host *gag* diversity accumulated at a rate of 0.3% per year in a 700bp fragment spanning p17 to p24, between an early time point no later than 9.8 months following seroconversion and a second time point at least 2.0 years later. The estimate from the previous study used a slightly different methodology of tMRCA estimation, a shorter *gag* fragment, and involved non-B subtypes. However, it seemed probable that P1 had been infected with at least one epidemiologically unlinked strain of *gag*.

5.4.2.7 Recombination analysis in gag

Another way to visually represent the degree of genetic distance for the *gag* fragment within P1 was to make a Highlighter plot (Figure 5.5), in which sequences sampled from P2 up to day 235 were included for comparison.

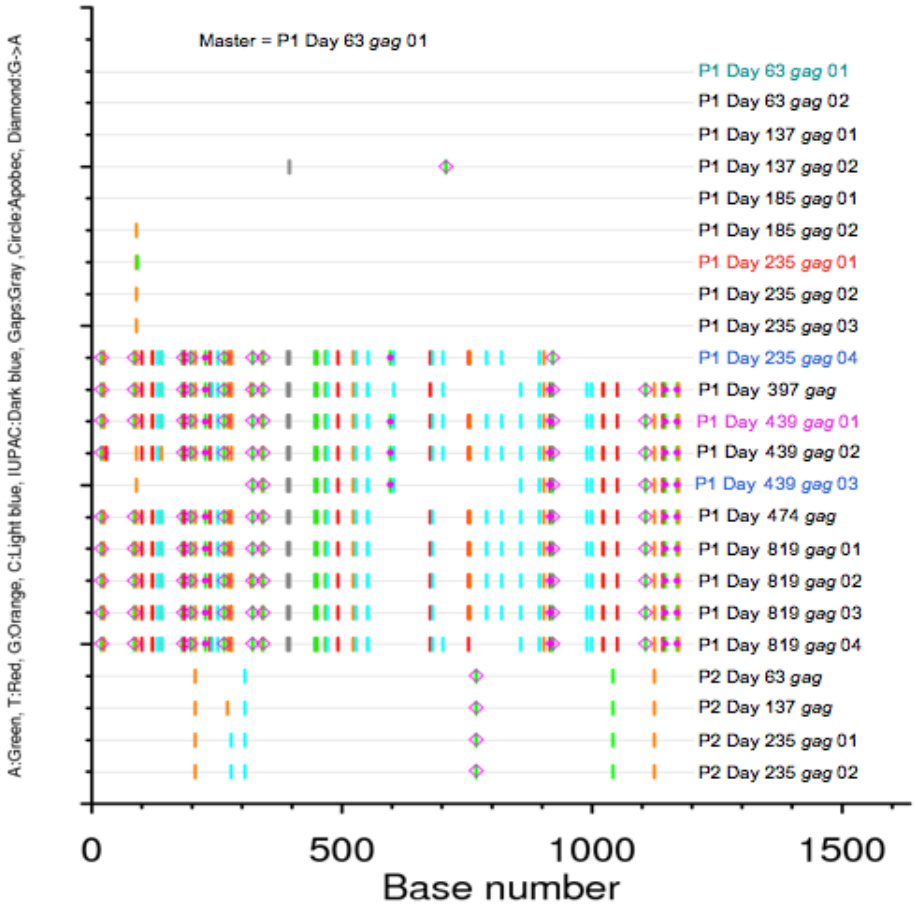


***Figure 5.5 Highlighter Plot of* gag.** *The names of suspected recombinant sequences are highlighted in blue font. The name most divergent sequence, used as the surrogate reference for the invading sequence, is highlighted in purple. The names of example cluster 1 sequences from days 63 and 235 are highlighted in green and red font, respectively.*

Recombinant sequences were initially inferred using a distance-based, query-versus reference method of screening, as implemented in RIP v3.0 (data not shown), with a p-value less than 0.05. The findings in RIP were confirmed using distance and phylogenetic (Kimura 2-parameter +NJ) methods as implemented in SimPlot

v3.5.1 (Figures 5.6 and 5.7).

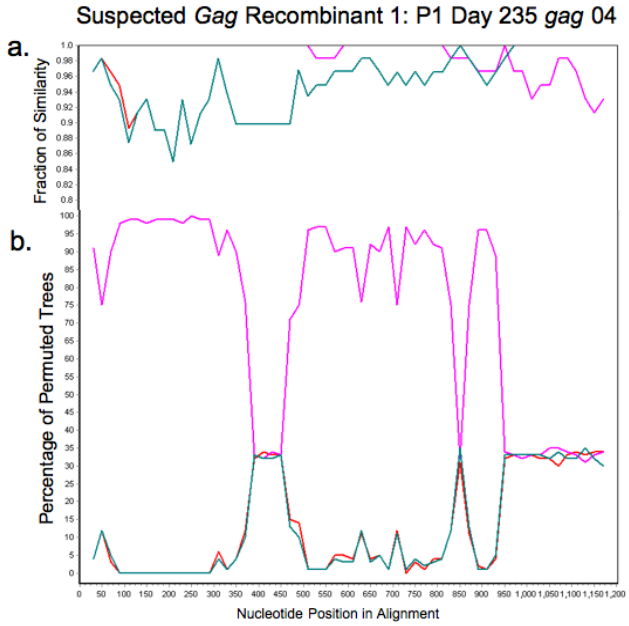


Figure 5.6 Query-versus-reference plots for the suspected recombinant sequence P1 Day 235 gag 04 of a. the fraction of sequence similarity (out of 1.0) and b. the percentage of permuted trees that support clustering with the reference sequence. *The green line represents the reference sequence P1 Day 63 gag 01. The red line represents the reference sequence P1 Day 235 gag 01. The purple line represents the reference sequence P1 Day 439 gag 01.*

The approximate location of breakpoints in the suspected recombinant sequences was confirmed by a third method, the mosaic detection algorithm, as implemented in 3Seq (Boni et al. 2007; Hogan and Siegmund 1986).

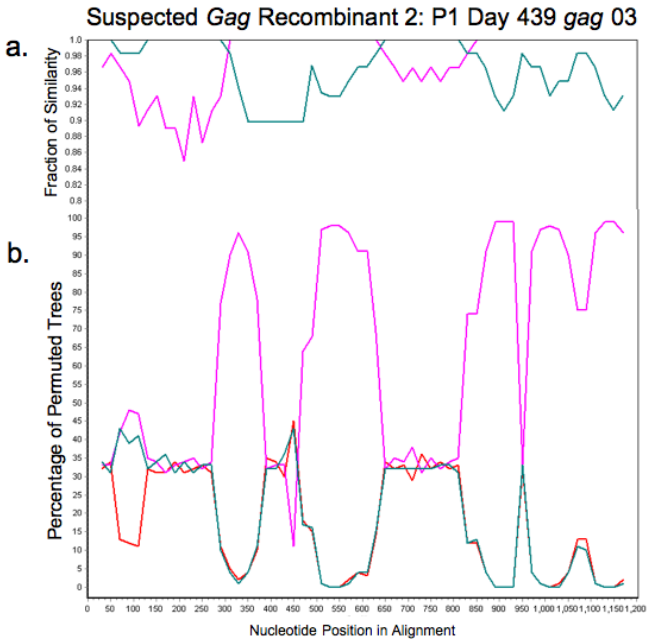


Figure 5.7 Query-versus-reference plots for the suspected recombinant sequence P1 Day 439 gag 03 of a. the fraction of sequence similarity (out of 1.0) and b. the percentage of permuted trees that support clustering with the reference sequence. *The green line represents the reference sequence P1 Day 63 gag 01. The red line represents the reference sequence P1 Day 235 gag 01. The purple line represents the reference sequence P1 Day 439 gag 01.*

Breakpoint patterns were different for each suspected recombinant sequence. The plots for the percentage of permutated trees were not conclusive for lineage switching for either suspected recombinant, because the bootstrap support for identity with either sequence was low in the regions identified to be most like the day 63 in the percentage similarity plot. In addition, the existence of breakpoints did not prove that the sequences sampled by population amplification were true recombinants, but interpreted alongside the other sequences sampled at days 235 and 439, there appeared to be more than one strain circulating at those time-points.

5.4.3 Immunological analysis

5.4.3.1 Class I HLA-mismatched mutation analysis in the gag protein sequence

I initially looked for mutations in the *gag* sequences from P1 and P2 that might indicate transmission of class I HLA-mismatched mutations from the donors that may affect replication capacity (Brockman et al. 2010; Miura et al. 2009a). Firstly, I compared the Class I HLA type of P1 and P2 with the clinical data to see if there was evidence of selection. P1 possessed HLA-A*0201, A*2402, HLA-B*1402, B*3543, Cw*0102, Cw*0802; P2 possessed HLA-A*0101, A*2901, B*0801, B*5001, Cw*0602, Cw*0701. Neither participant possessed HLA types that are strongly associated with protection from progression in chronic infection (X. Gao et al. 2001; Kiepiela et al. 2004). Secondly, I did not find any gag protein mutations in either P1 or P2 that were known to be associated both elite control of viral load *in vivo* and reduced replication rates *in vitro* (Brockman et al. 2010; Miura et al. 2009a; Miura et al. 2009b). However, P1 acquired a mismatched HLA-associated mutation in the suspected invading sequence within the epitope TSTLQEQIGW (TW10). The TW10 mutation TSTLQEQIAW (G248A) is associated with hosts carrying the HLA-B*5801 allele (Martinez-Picado et al. 2006), but neither P1 nor P2 carried this allele. Both P1 and P2 were infected with the wild-type TW10 mutation before week 235, and the sampled virus in P2 also had the wild type epitope at week 235. However, the G248A mutation was found in the most divergent sequences in P1 from week 235, with the exception of one the suspected recombinant sequence P1 day 235 *gag* 04.

5.4.3.2 Acquisition of a Class I HLA-matched escape mutation by P1

I then looked at the SLYNTVATL (SL9) epitope in gag, which is known to be restricted by HLA-A*0201, because P1 carried the HLA-A*0201 allele while P2 did not. Specific SL9 mutations are also

associated with lower viral loads *in vivo* during chronic infection (Iversen et al. 2006; Karlsson et al. 2007). At day 63 and day 137, all sequences from both P1 and P2 contained the wild-type SL9 epitope. Most sequences from day 235 in P1 contained the known SLFNTVATL (Y79F) mutation for escape from HLA-A*0201 (Goulder et al. 1997; Harrer et al. 1998; Iversen et al. 2006; Jamieson et al. 2003) instead of the wild-type SL9 epitope. Only the two recombinant sequences, detected at days 235 and 474, contained the wild-type SL9 epitope.

5.4.3.3 CTL response to SL9 or variants

Despite carrying the HLA-A*0201 allele, P1 did not mount a detectable CTL response on ELISpot to either the wild-type SL9 or previously described Y79F-containing mutants at days 63, 235 or 474 (Chapter 8). Although generation of the escape mutation within P1 could not be excluded, the appearance of the Y79F mutation would be consistent with invasion by an epidemiologically unlinked infection already carrying the escape mutation at SL9.

5.4.3.4 All HLA matched CTL responses to overlapping gag peptides in P1

I then looked at all the peptides for which P1 made a response on γ-interferon ELISpot and compared the sequences in P1 and P2 (Table 5.2). Notably, P1 did not make any ELISpot responses against the available sequences at day 474. I then compared all homologous sequences present in each cluster in P1. At least one mutation was found in all four of the overlapping peptide homologous sequences on P1 Cluster 2, except in recombinant sequences. P1 and P2 did not make ELISpot responses against the same OLP and did not contain the same mutations, except where the P2 day 235 sequence and P1

Cluster 2 contained the same K-to-R mutation at position six (HXB2 gag AA 30) in the overlapping peptide G7. These findings may be consistent with invasion by mutants carrying escape mutations but whether or not the mutations detected are true escape mutations is unknown.

Table 5.2 ELISpot responses for P1 in the regions sequenced. The amino acid sequences for each SPARTAC ELISpot peptide code are given. The HXB2 amino acid location (AA) is given and mutations away from the SPARTAC ELISpot peptide sequence are highlighted in red, bold and underlined font. The ELISpot response units are spot firming units (SFU) per million cells. OLP = overlapping peptide, OP = optimal peptides. Date: P1 Response gives the day (e.g. d63 = day 63) when a positive response was detected and its magnitude (minus the background and 50 SFU per million cells). The matched Class I HLA is given for the optimal and variant peptides. Minor variants seen only in inferred recombinants in P1 Cluster 2 are shown in black bold.

Epitope +/- HLA	HXB2 AA Location	Day: P1 Response	Sequence
OLP			
gag 2	gag 5-19 (p17)	d63: 160	SPARTAC: ASVLSGGELDRWEKI P1 d63: ASVLSGGELDRWEKI P2 d63: ASVLSGGELDRWEKI P1 Cluster 1: ASVLSGGELDRWEKI P1 Cluster 2: AS**I**LSGGELDRWEKI AS**I**LS**G**GELDRWEKI **ASVLSGGELDRWEKI** P2 d235: ASVLSGGELDRWEKI
gag 7	gag 25-39 (p17)	d235: 280	SPARTAC: GKKKYKLKHIVWASR P1 d63: GKKKYKLKHIVWASR P2 d63: GKKKYKLKHIVWASR P1 Cluster 1: GKKKYKLKHIVWASR P1 Cluster 2: GKKKY**R**LKHIVWASR P2 d235: GKKKY**R**LKH**V**VWASR
gag 34	gag 133-147 (p17)	d63: 60	SPARTAC: PIVKQNLQGQMVHQAI P1 d63: PIV-QNLQGQMVHQAI P2 d63: PIV-QNLQGQMVHQAI P1 Cluster 1: PIV-QNLQGQMVHQAI P1 Cluster 2: PIV-QNLQGQMVHQ**P**I P2 d235: PIV-QNLQGQMVHQAI
Epitope +/- HLA	HXB2 AA Location	Day: P1 Response	Sequence
gag 75	gag 297-311 (p24)	d63: 200 d235: 100	SPARTAC: VDRFYKTLRAEQASQ P1 d63: VDRFYKTLRAEQASQ P2 d63: VDRFYKTLRAEQA**T**Q

			P1 Cluster 1:	VDRFYKTLRAEQASQ
			P1 Cluster 2:	VDRFYKTLRAEQASQ
				VDRFYKTLRAEQASQ
			P2 d235:	VDRFYKTLRAEQATQ
OP				
nef 3 (HLA-A*0201	nef 83-91	d63: 120	SPARTAC:	AALDLSHFL
			P1 d63:	GALDLSHFL
			P2 d63:	AALDMSHFL
			P1 d235:	GALDLSHFL
			P1 d474:	GWLDLSHFL
			P1 d819:	GWNDLSHFL
			P2 d187:	AALDMSHFL
env 24 (HLA-A*0201)	env 375-383 (C3 region of gp120)	d63: 60	SPARTAC:	SFNCGGEFF
			P1 d63:	SFNCGGEFF
			P2 d63:	SFNCGGEFF
			P1 d235:	SFNCGGEFF
			P2 d474:	SFNCGGEFF
			P1 d819:	SFNCGGEFF
			P2 d235:	SFNCGGEFF

5.4.3.4 HLA-matched CTL responses to optimal peptides in P1

P1 made CTL responses against two optimal peptides, one of which was in nef and the other was in env gp120 (Table 5.2). Both had been previously associated with HLA-A*0201, and P1 carried this common allele while P2 did not. The A79G mutation in the nef optimal peptide had previously been associated with HLA-A*0201 in subtype B (Karlsson et al. 2007). The A79G nef mutation was present in P1 at day 63 and so likely reflected early within-host targeting of the original strain of nef (Brumme et al. 2008b). In addition, since the HLA-matched SL9 mutation had appeared in P1 in the absence of a detectable immune response, it was possible that an HLA-A*0201 epidemiologically unlinked donor had also mutation, because it was present in all the sequences from the second (N2) cluster (days 235,

474 and 819) in P1. However, given that P1 already had the mutation, so possible escape did not explain rapid superinfection of P1 in nef by day 235. As for the env optimal peptide, despite the detection of an env response at day 63, no change was seen in the sequence of the peptide, although mutations were seen either side of the epitope, indicating that this peptide may be functionally conserved (C. Otto et al. 2003). The relevance of the env protein mutations to the invading sequences seen in nef and gag were unknown but were explored for Chapter 6. In summary, superinfection due to fitness effects of the mutations seen may not be attributable simply to evasion of detectable immune responses.

5.5 Discussion

5.5.1 Identification of an epidemiologically unlinked, invading sequence

5.5.1.1 Different patterns of clustering in the sampled genes

The separate clustering of the P1 *nef* sequences from day 235 was consistent with an epidemiologically unlinked strain replacing the original infection in that gene. For *env*, no distinct second cluster was seen, with all the P1 sequences continuing to cluster together. In gag the two clusters were not separated by epidemiologically unlinked sequences, although the second cluster of P1 sequences from day 235 was less similar to the day 63 P1 sequences than the day 63 P1 sequences were to the P2 sequences at days 137 and 235. By both distance-based and ML estimation, the same sequence from day 434 was the most divergent from the sampled sequences in P1 at day 63.

5.5.1.2 Use of a surrogate reference for the invading strain

The identity of the invading sequence was inferred based on sequence divergence because the epidemiologically unlinked donor was not sampled. Therefore, the most divergent sequence at day 434 was used as a surrogate for the parental sequence of the invading strain in *gag*. Had both the supposed original and new donors been traced and sampled, definitively proving the direction of transmission may not have been possible (Pillay et al. 2007).

5.5.1.3 Relevance of recombinant sequences sampled by population amplification

Since population amplification sequences had been used in the analysis, recombinants may have been generated artificially by taking the consensus and because of PCR-induced recombination

(Jordan et al. 2010). However, since the main purpose was to detect the possibility of multiple-infection, determining the true *in vivo* haplotype for the recombinants was less important than the detection of different variants in homologous sequences at the same time point.

5.5.2 Immunological implications of multiple-infection

5.5.2.1 Screening for HLA mismatched mutations in the gag protein that are associated with reduced replication capacity in vitro

It had been hypothesized that the immune response favours replacement of the original strain with a superinfecting strain, and that natural selection may produce a recombinant superinfecting strain with advantageous escape mutations (Altfeld et al. 2002; Streeck et al. 2008). I did not detect any HLA-associated mutations in any of the gag protein sequences from P1 or P2 that had previously been associated with reduced replication capacity *in vitro* (Brockman et al. 2010; Miura et al. 2009a; Miura et al. 2009b). I considered Class I HLA-mismatched mutations in both P1 and P2 because donor-related HLA footprints could theoretically affect initial replication capacity prior to reversion (Duda et al. 2009; Frater et al. 2006; Matthews et al. 2008). P1 acquired an HLA-mismatched mutation (G248A) in the TW10 epitope at day 235 that was only found in the second P1 phylogenetic cluster and was not present in P2. The G248A mutation had not been inferred to have a fitness-cost *in vivo* or *in vitro*, and has been shown to persist upon transmission to an HLA-mismatched host (Martinez-Picado et al. 2006). However, it seemed unlikely that the G248A mutation should arise in P1 or P2 and so it would be consistent with invasion by an epidemiologically unlinked strain.

5.5.2.2 Acquisition of an HLA-matched mutation

I then looked at the gag epitope SL9 because this peptide is immunogenic and mutations in this peptide sequence are common as is the restricting allele, HLA-A*0201, in European populations (Goulder et al. 1997; Iversen et al. 2006; Kan-Mitchell et al. 2004; Karlsson et al. 2007). Particular mutations in SL9 are also associated with reduced viral load *in vivo* during chronic infection (Iversen et al. 2006). Interestingly, a known HLA-A*0201 associated escape mutant, Y79F (Goulder et al. 1997; Harrer et al. 1998; Iversen et al. 2006; Jamieson et al. 2003), was found in P1 from day 235. I did not find the Y79F mutation in sequences that clustered with the day 63 sequences in P1 and P2. The Y79F mutation likely leads to diminished, but not complete loss of, binding affinity between the SL9 peptide and the HLA molecule (Iversen et al. 2006; Jamieson et al. 2003).

5.5.2.3 Potential advantage of acquiring HLA-matched escape mutations through multiple-infection

It was also possible that acquisition of the Y79F mutation by either within-host adaptation or multiple-infection might become advantageous to the virus in an HLA-A*0201 positive host and may even contribute to population-level adaptation to a common HLA allele (Iversen et al. 2006; Karlsson et al. 2007). Human hosts that carry the HLA-A*0201 allele may be able to weakly recognize the Y79F escape mutant but the response is generally less than to the wild-type SL9 (Harrer et al. 1998; Iversen et al. 2006; Jamieson et al. 2003).

5.5.2.4 Lack of a detectable immune response specific for SL9 in P1

However, without evidence of a specific immune response in P1

during the first year of infection, it seemed unlikely that mutation had occurred due to specific targeting of the SL9 epitope in P1. Although P1 carried the correct restricting HLA allele, there was no evidence of a significant immune response to either the wild-type SL9 peptide or any of the tested SL9 variant peptides containing the Y79F mutation. The lack of HLA-A*0201 ELISpot responses to the SL9 wild-type infection at day 63 may reflect insensitivity in the assay or the observation that SL9-directed responses and mutations are uncommon during the first year of HIV-1 infection (Duda et al. 2009; Kan-Mitchell et al. 2004). Escape mutations in SL9 appear to play a more significant role during the chronic phase of infection (Iversen et al. 2006; Karlsson et al. 2007). It was also possible that an invading SL9 escape mutant might 'hitch-hike' within a generally fitter gag protein if recombination within the *gag* gene were limited by reduced sequence identity (Nikolaitchik et al. 2011).

5.5.3 Comparison with other cases of intra-subtype multiple infection or superinfection

5.5.3.1 Detection of multiple-infection or superinfection within the same subtype

Other case studies of intra-subtype multiple-infection or superinfection have been reported in the literature. Multiple-infection is generally inferred by the presence of superinfecting strain but co-circulation of more than one variant was detected in only two previous cases. A summary of previously published case studies is given in Table 5.3. An additional potential case of subtype B multiple-infection was found in an IVDU who had not progressed to AIDS after 18 years, but the authors had not been able to exclude co-infection by sampling before the appearance of dual infection (Casado et al. 2007).

5.5.3.2 Fitness implications of detecting multiple-infection

The detection of co-circulation is important for determining whether potentially fitness-altering mutations are associated with the invading strain or whether the mutation evolved within the host. One of the studies that detected rather than inferred multiple infection investigated multiple infection with drug-sensitive and drug-resistant strains, with the latter forming the original infection (Koelsch et al. 2003). Only one study measure *in vitro* replication rate (van der Kuyl et al. 2010), and this was for the *pol protease-RT* region. The authors concluded that the invading strain had a faster

Table 5.3 Other published cases of intra-subtype multiple-infection or superinfection. The number of participants (N) infected with more than one intra-subtype strain, the subtype and whether multiple-infection or superinfection was detected. Only participants infected with the same subtype have been included.

Reference(s)	N	RF	Sub-type	Gene	Multiple infection	Super-infection
(van der Kuyl et al. 2010)	2	MSM	B	*gag*, *env*	Yes	Yes
(Rachinger et al. 2008; Rachinger et al. 2011a)	1	MSM	B	*env*	Inferred	Yes
(Streeck et al. 2008)	1	MSM	B	*gag*, *env*	Inferred	Yes
(D. M. Smith et al. 2005b)	1	MSM	B	*pol*	Inferred	Yes
(Koelsch et al. 2003)	1	MSM	B	*gag-pol*, *env*	Yes	Yes
(Altfeld et al. 2002)	1	MSM	B	*gag*	Inferred	Yes

replication rate, but the implications of multiple-infection for the relative replication rate of *gag* proteins remains unknown.

5.5.3.3 Drawbacks of analysing single population-amplification sequences where recombination is inferred

The authors of previous studies have analysed population amplification sequences, which means that it is unclear whether the genetic sequences represents *in vitro* recombination, *in silico* recombination from construction of the consensus sequence, or true *in vivo* recombination (Streeck et al. 2008). To demonstrate co-circulation of multiple variants in P1, rather than superinfection with a single recombinant strain, I to clone the *gag* amplicons and perform SGA on *env* (Chapter 6).

5.6 Conclusion

5.6.1 Key Findings

I have demonstrated that sequences within P1 sampled after day 235 in *nef* and *gag* are consistent with invasion by an epidemiologically unlinked strain of HIV-1 subtype B following exposure to an additional person after day 63. Invasion by the epidemiologically unlinked strains was not observed in P2. I could not attribute escape from the immune system in the gag protein of P1 to the epidemiologically unlinked invading strain. However, a known HLA-matched escape mutant in P1 appeared co-incident with the invading strain cluster sequences despite the absence of a detectable immune response against the SL9 wild-type initially detected in both P1 and P2. I also noted the appearance of an HLA-mismatched mutation in the gag protein of P1 at day 235, as well as the appearance of escape mutations to detected CTL responses against some peptides in P1 that were found only in the inferred invading strain.

5.6.2 Further work to be done

5.6.2.1 Assay for gag-protease *determined replication capacity*

A key question remains as to whether or not the replacement of one strain by another in *gag* represents a change in the replication rate of the new strain. Parallel *in vitro* replication rate assays have been developed that specifically test the *gag-protease* region of isolated strain (Brockman et al. 2010; Miura et al. 2009a; Miura et al. 2009b) and a logical next step from the work presented in this chapter would be to apply such an assay to clones representing the original and invading sequences detected in P1. Furthermore, a comparison

could be made with the baseline sequence in P2, as P2 did not appear to have been invaded by the invading strain in P1, despite being exposed to his partner P1 as well as the new donor. If there is no difference in replication fitness between the original and invading strains in P1, or if the invading strain is less fit, then the superinfection of P1 by the invading strain may be attributed to the immune response rather than competition.

5.6.2.2 Additional nef *amplification*

Another additional piece of work would be to attempt amplification of *nef* at time-points from day 137 to day 235 in P1 and P2 to see if multiple infection rather than superinfection can be detected in that gene. This would be especially interesting, as P1 made early CTL responses detected against an epitope in nef. Furthermore, nef is known to be an early target of CTL immune responses (Brumme et al. 2008b) and it would be interesting to know whether or not rapid superinfection by the invading strain after multiple-infection could be attributed to early immune selection pressure.

5.6.3 Concluding remarks

Multiple infection of a single host provides an opportunity to observe *in vivo* the relative fitness of two strains. Superinfection by a single dominant strain appears to be a gradual process, even in relatively conserved genes. Eventual superinfection may be attributable to a balance between replication rate and the ability to avoid an immune response.

Chapter 6 A qualitative diploid model of multiple-infection and superinfection

6.1 Overview

6.1.1 Brief summary of chapter and hypothesis

6.1.1.1 Overall hypothesis and background

The overall hypothesis of Chapter 6 is that HIV-1 behaves like a diploid (or pseudodiploid) organism, rather than as a classical haploid organism. If HIV-1 behaved like a classical haploid organism, as proposed by previous, widely-cited models of HIV-1 evolution within hosts (Boerlijst et al. 1996; E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009; Lemey et al. 2006; McLean and Nowak 1992; Nowak and May 1994), then one would expect complete replacement of the original strain by a fitter invading strain. By contrast, if HIV-1 behaved like a diploid (or pseudodiploid organism), one would expect that the invading strain would be able to invade the viral population within a host in a number of different ways depending on the circumstances of a particular scenario (S. P. Otto and Day 2007), including scenarios where the average relative fitness of a mixed recombinant state would determine whether or not the initial infection can be invaded by a particular exogenous strain .

6.1.1.2 Approach taken in this chapter

Initially I used an existing, classical model of diploid evolution as a simple, qualitative model for virus evolution with the innovation that, instead of considering the relative fitness of two alleles, I considered the average relative fitness of the initial and invading

strains. Thus, I applied my qualitative model in a novel way, by using concepts and techniques from classical population genetics and inclusive fitness theory to make predictions about the conditions under which HIV-1 multiple-infection or superinfection may occur. Having worked through the various, predicted scenarios inmy simple, qualitative model, I then sampled multiple *env* and *gag* sequences at various time-points in P1 to confirm the time-scale over which the invading strain took to appear and then for a new, recombinant strain to dominate. I demonstrated that sequences from multiple strains, representing both the initial infection as well as the invading strain, could simultaneously co-exist for several months in the same patient before eventual domination by a recombinant strain. I re-evaluated both the sequence and immunological data to see if any of the sampled sequences representing the initial or the invading strain may carry mutations that would be advantageous. I then applied my simple qualitative model to explain why recombinants that may more or less represent the initial or the invading strain may come to dominate despite no clear fitness advantage in terms of the detected immunological responses due to the average relative fitness of the mixed recombinant state and the very large initial frequency of the initially infecting strain versus that of the invading recombinant strain. Finally, I use my model to predict that detection of superinfection in HIV-1 is uncommon, not because the initial strain has a superior relative fitness, but that its very high initial frequency, plus the average relative fitness of a mixed recombinant state may be crucial to the successful invasion of the host viral population by an invading strain. This prediction is of significance because it may

explain why superinfection is uncommon despite the difficulty in eliciting consistent evidence of a protective immune response in most HIV-1 infected hosts, or most HIV-1 naïve but vaccinated hosts. Therefore, new approaches to preventing either the transmission to, or the infection of, new hosts by HIV-1 may require novel approaches. Such novel approaches may not simply rely on the generation of (partially) protective immunity at the mucosa but rather also aim to subvert viral mechanisms that are critical to HIV-1 invasion and adaptation to new hosts, including recombination.

6.2 Introduction

6.2.1 Overview

6.2.1.1 Overview of deficiencies in current models for explaining observations of multiple-infection

One of the most important questions facing researchers of an HIV-1 vaccine is to explain the apparent paradox between the ineffectiveness of the human immune response in preventing progression to AIDS and the apparent rarity of multiple-infection or superinfection with epidemiologically unlinked strains (Rachinger et al. 2011b). Existing models, based on the assumption of simple competition between two haploid genomes, do not allow for a recombinant viral state to exist (Murray 2002; S. P. Otto and Day 2007). By extension, these models of HIV-1 infection assume that the dominant viral strain represents the “fittest strain” and that other strains cannot invade the host due to being less “fit” than the dominant strain (Murray 2002; S. P. Otto and Day 2007). Some authors have hypothesized that the immune response to the initial infecting strain can prevent invasion by an epidemiologically unlinked strain while being incapable of eradicating the initial strain, since the initial strain has had more opportunity to adapt to the host and its immune response (Altfeld et al. 2002; Rachinger et al. 2008; Rachinger et al. 2011a; Streeck et al. 2008). However, no absolutely preventative immune mechanism has yet been detected in infected human hosts and there is as yet no explanation for why some hosts become multiply infected while other, similarly exposed hosts do not (Altfeld et al. 2002; Rachinger et al. 2008; Rachinger et al. 2011a; Streeck et al. 2008).

6.2.1.2 Haploid models do not account for inferred recombination between the initial and invading strains in multiple-infection

Although existing models do not currently allow for a recombinant state, recombination between the original and invading strains has been detected (McCutchan et al. 2005; Streeck et al. 2008). Previous authors have speculated that these recombinants may have a selective edge in evading host immune responses, and thus displace both the original and invading strain (McCutchan et al. 2005; Streeck et al. 2008). However, while these hypotheses acknowledge the existence of recombination, there has been no modification of the underlying assumption of haploidy to explain the presence of recombination.

6.2.1.3 Overview of the contributions in this chapter

In this chapter, I introduce a simple qualitative model for competition between the initial and invading strains of HIV-1 that assumes a diploid genome and is thus an important starting point for explaining not only the presence, but also the vital role of, recombination in explaining the apparent rarity of multiple-infection in the absence of a preventative immune response. The data presented in this chapter support the existence of recombination between the initial and invading strains, and hence the need to use diploid models. The model presented aids in the interpretation of the data by providing an explanation for invasion by an epidemiologically unlinked strain in the presence of significant recombination with the initial strain without the need to include mechanisms that are not detected experimentally.

6.2.2 Conflicting data predictions in previous approaches

Previous interpretations of existing data have been confusing not

only due to experimental limitations but also because of the limitations of theoretical predictions based on haploid models of competition between strains. Experimental limitations could be overcome to some extent if an internally consistent model were applied to explain the observable data. In HIV-1, multiple-infection, superinfection and recombination have previously been modelled using extensions of quasispecies theory (Boerlijst et al. 1996; Nowak and May 1994) or coalescent-based models (Arenas and Posada 2010; D. J. Wilson and McVean 2006). However, the critical flaw with the quasispecies approach is that HIV-1 does not satisfy central assumptions of quasispecies models, such as the assumption of a haploid organism (Boerlijst et al. 1996; Holmes 2009; Lemey et al. 2009). Furthermore, the problem with using an alternative coalescent-based approach is that many of the mechanistic parameter values are not known with any certainty when quantifying estimates of the effect of selection, mutation, recombination, migration and population dynamics within HIV-1 infected hosts (Perelson et al. 1996; Stafford et al. 2000). New modelling is therefore required to target experimental work with limited datasets.

6.2.3 Qualitative approach to interpreting experimental data

6.2.3.1 The need for a simple qualitative diploid model

To overcome the difficulty of data limitations, I adopted a simple qualitative model to begin understanding the effect of having a diploid genome on the ability of HIV-1 to cause multiple-infection or superinfection. The simple, qualitative diploid model was important in its own right and as a starting point for future development of more complicated models of multiple-infection in the presence of recombination. Initially, I developed a novel application of an existing diploid model in classical population genetics, to account for pseudo-diploid replication characteristics specific to HIV-1 that deviate from the Mendelian, allele-based interpretation of such classical models. (Haldane 1932; Leigh 1990; S. P. Otto and Day 2007)

6.2.3.2 Targeting experimental work based upon a novel interpretation of a classical diploid model

I then set out to demonstrate experimentally that recombination was in fact occurring within the *gag* and *env* genes of viral strains infecting the human host, P1, whom I have shown to be multiply infected in Chapter 5. I achieved this by using different experimental methods appropriate for the gene under study and sampling from multiple time points. I then set out to try and experimentally detect if recombination could be attributed to some partially effective immune response against the initially infecting strain that would confer some advantage on the invading strain or a recombinant.

6.2.3.3 Interpreting the experimental data based upon a novel interpretation of a classical diploid model

Where an effective, or partially effective, immune response could

not be detected, I turned to my simple qualitative diploid model to see if I could explain my experimental recombination data without the need for any immune response. By extension, I informed my understanding of why multiple infection and superinfection may be uncommonly detected in the absence of a protective immune response. In the discussion, I consider the relevance of the model and the experimental data from my study of multiple-infection and superinfection to the different scenario of transmission of, and co-infection with, more than one variant during primary HIV-1 infection, as this has implications for experiments that set out to try and characterize transmissible diversity and infer characteristics of the initial immune response. Lastly, I discuss the implications of my simple qualitative model for potential strategies that may block HIV-1 infection in the absence of an ability to elicit even a partially effective immune response. Specifically, blocking the ability of the virus to recombine may impair the ability of HIV-1 to invade and adapt to infection-naïve as well as infected human hosts.

6.3 Model

6.3.1 Definitions and Assumptions

6.3.1.1 Basic assumptions of the simple model

The complex system of HIV-1 evolution is simplified to a discrete, two-variant model of diploid viruses with non-overlapping generations lasting approximately 2 days, defined as *t*. The mathematics for the simple diploid model of HIV-1 is the classical Hardy-Weinberg law (Crow and Kimura 1965; Leigh 1990; S. P. Otto and Day 2007), except that the model is applied to genetic subunits of that undergo recombination, rather than alleles undergoing sexual reproduction as in Mendelian genetics. Each genetic subunit has a phylogenetic different history, so that the genetic subunit from the original strain belongs to phylogeny *A* and the genetic subunit from the invading strain belongs to phylogeny *B*. Latent infection, overlapping generations and stochastic effects are not included in the simple model of HIV-1 I present here.

*6.3.1.2 The original strain (**AA**)*

The model assumes that the host population is infected with an original strain of diploid virus, called pure strain *AA*, which has a single phylogeny and initially exists at very high frequency within the host. All diploid virions of the original strain contain two RNA strands, with each being a pure sequence of the same phylogeny, *A*. These *AA* virions are reverse transcribed into haploid proviruses that contain genetic material only from phylogeny *A*.

*6.3.1.3 The epidemiologically unlinked invading strain (**BB**)*

I then assume that the host is invaded by an epidemiologically unlinked strain, called pure strain *BB*, which has its own phylogenetic history, and initially exists at a very low relative

frequency within the human host. All diploid virions of the epidemiologically unlinked invading strain contain a copy of the genes on two RNA strands, with each strand bearing a sequence of phylogeny *B*. These *BB* virions are reverse-transcribed into proviruses that contain genetic material only from phylogeny *B*.

6.3.1.4 Relative frequency of genetic sequence for phylogenies **A** *and* **B**

In HIV-1, the quantities of interest are the relative frequencies of units of genetic sequence from each phylogeny over time. The relative frequency of genetic sequences of phylogeny *A* is termed $p(t)$ and the relative frequency of sequence of phylogeny *B* is termed $q(t)$. Since we assume there are only two species in this simple model, $p(t) + q(t) = 1$.

6.3.2 The discrete deterministic model

6.3.2.1 Initial conditions

I assume that the initial relative frequency of the original strain is very close to 1 and the initial relative frequency of the invading strain is very close to 0. The true initial relative frequencies in HIV-1 infection are unknown but this assumption is valid as the original infection probably greatly outnumbers any strain in infectious bodily fluids during sexual transmission.

6.3.2.2 Relative fitness

The change in the relative frequency of sequences of phylogeny *A* or *B* over time in a simple diploid model of selection depends upon the relative fitness, *W*, of each strain in which these sequences are found (Leigh 1990; S. P. Otto and Day 2007). The relative fitness is the fitness consequence for the individual genetic subunit of selection in the diploid virion when it recombines with another subunit to

form a provirus that is more or less like either pure state (*AA* or *BB*) or the mixed recombinant state (*AB*). The sequences are each found in the respective pure strains as well as to varying degrees in recombinant strains. The true fitness values are unknown so the aim of the model is to provide a qualitative prediction.

6.3.2.3 Simplifying assumptions for the classic, two-strains model

In order to apply techniques for a classic, two-strains model of diploid selection, I have simplified the complexity of HIV-1 infection by assuming that strains are either more like the pure original or invading strains or are more like the middle, mixed-recombinant state. The relative fitness term for the pure *AA* strain is W_{AA} and the relative fitness term for the pure *BB* strain is W_{BB}. I assume 'pure' sequences of phylogeny *A* or *B* have relative fitnesses W_{AA} and W_{BB} which are an average effect of a virus recombining with sequences to form proviruses that are relatively more like either pure state rather than the recombinant state. The relative fitness for recombinant viruses (W_{AB}), does not denote the relative fitness of a specific recombinant strain in this model but to relative fitness when a gene recombines to form a provirus that is relatively more like the mixed recombinant state *AB*. The relative fitness of the mixed recombinant state is an average effect of the fitness consequences of recombination for an individual gene, as some recombination events will be lethal and the rest will result in different degrees of relative fitness.

6.3.2.4 Mathematical equations

The mathematical equations presented below are identical to the classical Hardy-Weinberg law as described by Leigh as well as Otto and Day (Leigh 1990; S. P. Otto and Day 2007), using the latter's notation. The recursion equation for the change in the relative frequency of sequences of phylogeny A over time, is (S. P. Otto and Day 2007) given by:

$$p(t+1) = p(t)^2 \frac{W_{AA}}{\overline{W}} + \frac{1}{2}\left(2p(t)q(t)\frac{W_{AB}}{\overline{W}}\right). \quad (1)$$

and for phylogeny B (S. P. Otto and Day 2007):

$$q(t+1) = q(t)^2 \frac{W_{BB}}{\overline{W}} + \frac{1}{2}\left(2p(t)q(t)\frac{W_{AB}}{\overline{W}}\right). \quad (2)$$

where the value $\overline{W}$ is called the mean population fitness and is defined as the sum of all phylogeny frequencies following selection, or (S. P. Otto and Day 2007):

$$\overline{W} = p(t)^2 W_{AA} + 2p(t)q(t)W_{AB} + q(t)^2 W_{BB}. \quad (3)$$

This value for $\overline{W}$ is used in equations (1) and (2) above give the phylogeny relative fitness terms (S. P. Otto and Day 2007).

6.3.2.5 Simulations

Simulations using this diploid model were coded and run in Matlab vR2010a (The Mathworks™, USA). The Matlab code can be found in the Appendix to this thesis. Simulations were run for 2000 generations, equivalent to approximately 4000 days (10.96 years) of HIV-1 infection, to ensure that equilibrium had been reached.

6.3.3 Discussion of simulation results in context of HIV-1

6.3.3.1 Overview

The model is classical with known results so more detailed discussions of the model can be found in standard texts (Crow and Kimura 1965; Leigh 1990; S. P. Otto and Day 2007). However I briefly discuss the implications of the model for predicting multiple-infection or superinfection in HIV-1 and provide illustrative simulations. As the aim of this simple model is to provide qualitative predictions of evolution, I use estimated parameter values that are illustrative of various scenarios under the qualitative model.

6.3.3.2 When the fitness of the original strain is greater than or equal to the relative fitnesses of the invading and recombinant strains

If the original strain of HIV-1 was fitter than or equally as fit as the invading and recombinant strains, then it always remained dominant and neither multiple-infection nor superinfection would occur (Figure 6.1). This may have occurred if the original strain were more optimally adapted for evading the immune response than other strains to which the host was exposed. It may also have occurred if the viruses from other hosts had a lower replication rate because of the cost of evading the donor immune response.

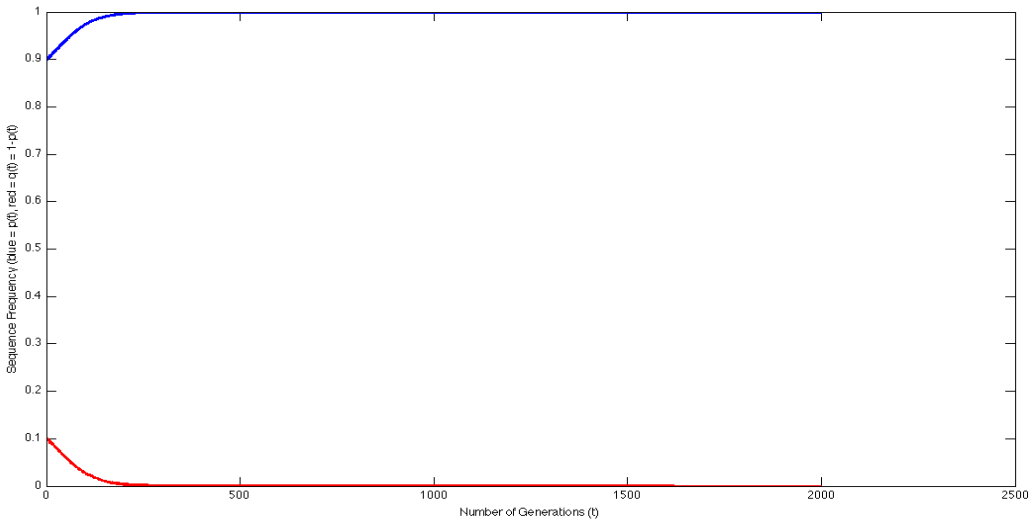


Figure 6.1 No multiple-infection or super-infection. *Relative frequencies of genetic subunits from phylogeny* A *(blue) and phylogeny* B *(red) are shown. The horizontal axis is time in generations (t) and the vertical access is relative frequency (0 to 1.0).*

6.3.3.3 When the fitness of the invading strain is greater than the original and recombinant strains.

If the invading strain had the highest fitness, the outcome was dependent upon the initial relative frequency of the original and invading strains as well as the relative fitness (Figures 6.1 and 6.2). If the original strain had a higher frequency advantage, as is probably the case in HIV-1 infection (Haase et al. 1996; Ho et al. 1995; Perelson et al. 1996), then the original strain would continue to dominate so long as its relative fitness was greater than that of the recombinant strain and it did not fall below a critical threshold relative to the relative fitness of the invading strain (Figure 6.1). The most important implication of this result is that even if the invading strain had the highest fitness, multiple-infection would not occur under these conditions. Superinfection would only occur (Figure 6.2) if: i) the initial frequency of the fitter invading strain were sufficiently high; ii) the relative fitness of the recombinant mixed state was higher than that of the original strain; or iii) if the relative

fitness of the invading strain was above a particular threshold relative to that of the original strains.

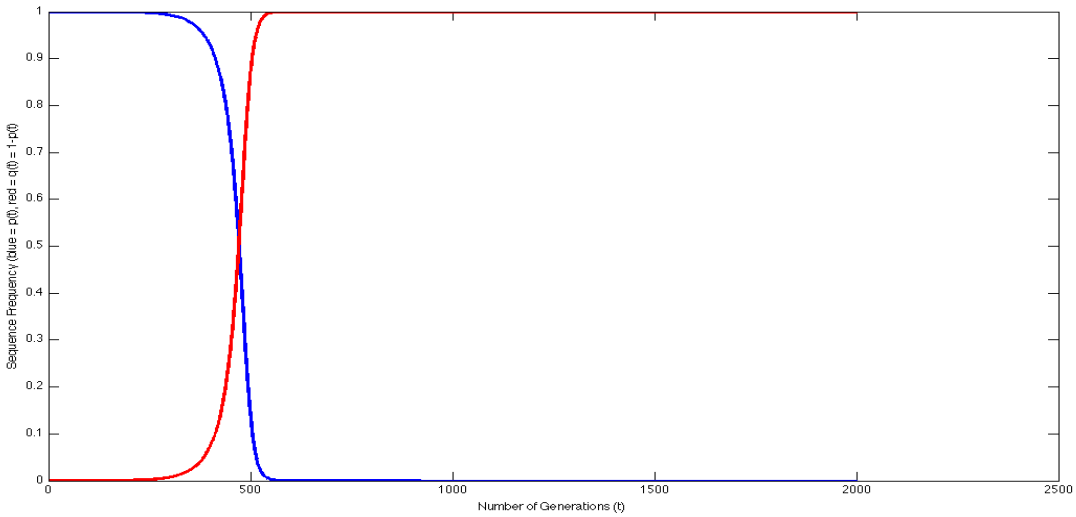


Figure 6.2 Superinfection by phylogeny B. *Relative frequencies of genetic subunits from phylogeny* A *(blue) and phylogeny* B *(red) are shown. The horizontal axis is time in generations (t) and the vertical access is relative frequency (0 to 1.0).*

6.3.3.4 When the recombinant state is on average fitter than either the pure original strain or the pure invading strain, and the pure original strain is more fit than the pure invading strain

In the scenario where the mixed recombinant state *AB* has both the highest relative fitness and when the relative fitness of pure original strain *AA* is greater than that of pure invading strain *BB*, stable multiple-infection will occur where the viral population will resemble phylogeny *A* more than phylogeny *B* (Figure 6.3).

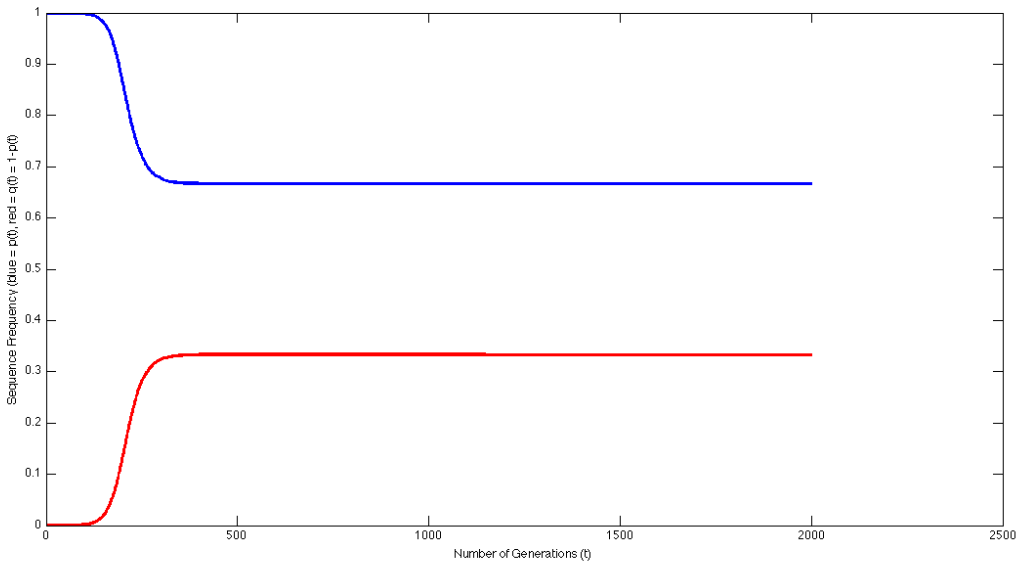


Figure 6.3 Multiple-infection when the population mostly resembles the original strain. *Frequencies of genetic subunits from phylogeny A (blue) and phylogeny* B *(red) are shown. The horizontal axis is time in generations (t) and the vertical axis is relative frequency (0 to 1.0).*

6.3.3.5 When the recombinant state is on average fitter than either the pure original strain or the pure invading strain, and the pure invading strain is more fit than the pure original strain

In the scenario where both the mixed recombinant state *AB* has the highest relative fitness and the pure invading strain's relative fitness *BB* is greater than that of pure original strain *AA*, stable multiple-infection will occur, where the viral population will resemble phylogeny *B* more than phylogeny *A* (Figure 6.4).

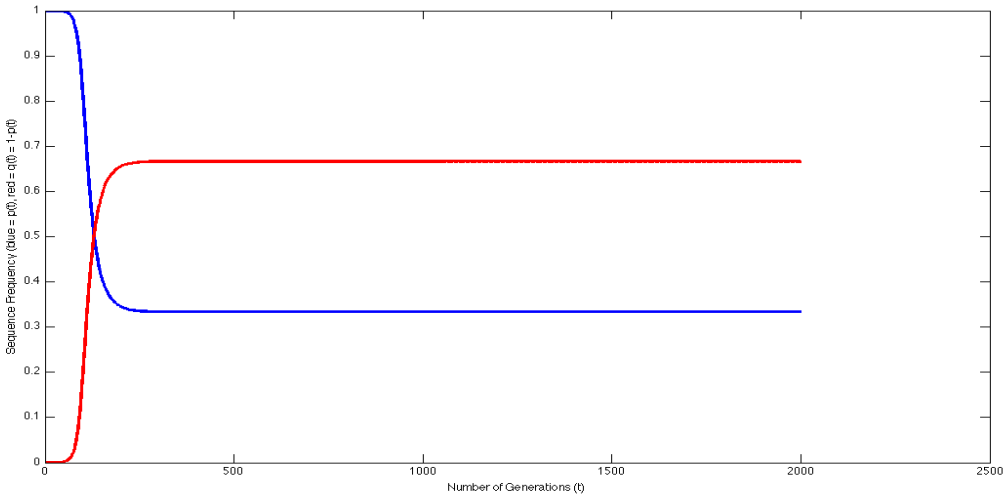


Figure 6.4 Multiple-infection when the population mostly resembles the invading strain. *Frequencies of genetic subunits from phylogeny A (blue) and phylogeny* B *(red) when the result is stable multiple-infection where the population more frequently resemble phylogeny* B*. The horizontal axis is time in generations (t) and the y access is relative frequency (0 to 1.0).*

6.3.4 Conclusions from the simple diploid model and simulations

The most important results from this simple diploid model are: (1) that a fitter invading strain may not necessarily succeed in causing multiple infection, and (2) that the relative fitness of the recombinant state is important for determining whether or not multiple infection to occur. Natural selection is known to have a lesser effect when diploid organisms undergo recombination and although this effect has been extensively analysed in Mendelian organisms (Crow and Kimura 1965; S. P. Otto and Day 2007), the effect of diploidy is infrequently considered in HIV-1 (Rambaut et al. 2004). While haploid models are useful for some aspects of HIV-1 evolution, the results here show that even very simple diploid models can inform our understanding of the underlying processes in multiple-infection and, by proxy, in co-infection.

6.4 Materials and Methods

6.4.1 Subjects

The subjects, P1 and P2, were previously described in Chapters 3 to 5. P1 and P2 presented to clinic together and were sampled for sequencing on the same day for all trial time points. However, samples for viral load and CD4+ leukocyte sampling were not necessarily taken on the same day for each patient.

6.4.2 Laboratory methods

Please see Chapters 3 and 6 for a full description of the laboratory protocols. For *gag*, population PCR amplicons from Chapter 5 were available for bacterial cloning. In chapter 5, population PCR in *gag* was attempted with multiple primer sets to try and confirm the dominant variant over time. The number of primer sets that were cloned for Chapter 6 at each time point is shown in Table 6.1.

Table 6.1 The number of primer sets represented in the clonal gag sampling from each time point for P1 and P2 . Cloning was unsuccessful after two attempts for the time points in red.

Sample (Days post-exposure)	P1 *gag* clones	P2 *gag* clones
63	2 primer sets	2 primer sets
137	1 primer set	1 primer set
185	2 primer sets	Not cloned
235	1 primer set	1 primer set
397	Not cloned	…
439	2 primers sets	…
474	1 primers set	…
819	3 primers sets	…

For *env*, single genome amplification was performed for the purpose of trying to demonstrate the existence of recombination. Two sets of primers were used for P1 Day 474 as it was challenging to

successfully amplify *env* amplicons at this time point using the primer set that had amplified the samples from days 63 and 235. Fewer time-points were attempted for env, due to the expense of single-genome amplification in terms of time and DNA template.

Table 6.2 The time points where* env *single genome amplification (SGA) was successful with a single primer set are indicated in black. The time point where SGA was successful with a single primer set but was attempted with two primer sets due to low sequence sample number is indicated in green. The time point where SGA was unsuccessful following attempts with two primer sets are indicated in red. Time points where SGA was not attempted are indicated in blue.

Sample (Days post-exposure)	P1 *env* amplicons	P2 *env* amplicons
63	See Chapter 3	See Chapter 3
137	RNA	RNA
185	RNA	RNA
235	RNA	RNA
397	RNA	…
439	RNA	…
474	RNA	…
819	RNA	…

6.4.3 Phylogenetic Analysis

6.4.3.1 Sequence alignment

Please see Chapters 3 and 6 for descriptions of sequence alignment methods. Only full-fragment length sequences were used for trees.

6.4.3.2 Screening for tree-likeness

If recombination had occurred, it would violate certain assumptions of the phylogenetic models used to reconstruct bifurcating trees. Therefore I initially screened the sequences for tree-likeness using uncorrected p values and the NeighborNet algorithm (Bryant and Moulton 2004; Lemey et al. 2009), as implemented in SplitsTree v4 (Huson 1998). Bootstrapping was performed with 1000 replicates.

6.4.3.3 Reconstruction of phylogenetic trees by ML estimation

Phylogenetic trees were initially constructed using the maximum likelihood (ML) method with PhyML v3.0 software (Guindon and Gascuel 2003) and visualized in FigTree v1.3.1 (Rambaut 2009). The chosen substitution model gave the highest likelihood when screened with jModelTest (Posada 2008, 2009). This chosen model was the GTR+Γ+I model (Büchen-Osmond). Bootstrapping was performed with 1000 replicates).

6.4.3.4 Reconstruction of phylogenetic trees using a Bayesian MCMC based approach

Phylogenetic analysis was repeated using a Bayesian MCMC based method, as implemented in MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). An unconstrained branch length (exponential) prior was used to avoid enforcing a molecular clock (Lemey et al. 2009). MrBayes v3.1.2 or v3.2 was run in duplicate for at least 50,000,000 sampling trees every 1,000 steps. Convergence was assessed with Tracer v1.5 (Rambaut and Drummond 2009) with all parameter estimates having ESSs of >300 (Lemey et al. 2009). The consensus tree for each gene, with posterior probabilities for branch support, was generated and visualized in FigTree v1.3.1.

6.4.4 Recombination Analysis

6.4.4.1 Visual screening for recombination breakpoints

Full nucleotide sequences for the fragments from all three genes were initially screened visually in Highlighter (LANL, US), using the baseline consensus for P1 as the master sequence for comparison. The sequences were then screened by a query-versus-reference approach with a distance-based method as implemented in RIP v3.0,

using the day 63 clonal sequence consensus for P1 as well as the most divergent clonal or SGA sequence, which for both *gag* and *env* was sampled at day 235. The window size used was 100 base pairs, which was the smallest allowable window size.

6.4.4.2 Phylogenetic-based screening

The query-versus-reference approach to phylogenetic-based recombination analysis was described in Chapter 5 using the most divergent clonal or SGA sequence, which for both *gag* and *env* was sampled at day 235. See Chapter 5 for a description of the use of SimPlot v3.5.1 (Lole et al. 1999) The window size used was 100bp and the bootstrapping used 1000 replicates.

6.4.4.3 Alternative algorithmic-based method of screening for recombination

The alternative algorithm-based approach for confirming breakpoints, as implemented in 3Seq, was described in Chapter 5. This method did not require a query-versus-reference approach and used a window size of 40bp.

6.4.5 Quantification of within-host diversification in terms of the tMRCA for the *env* sequences

I calibrated within-host evolution to the within-host tMRCA for the day 63 samples, as I had done in Chapter 4. The mean tMRCA prior estimate was 63 days with a standard deviation of one day. The demographic model used was the Bayesian Skyline plot with 10 groups. BEAST was run for at least 100,000,000 steps. Convergence was assessed in Tracer v1.5. Convergence was achieved when the ESS was >300 for all parameters. Strict and relaxed clock analyses were attempted. The substitution models tested were GTR, TN93 and TN93+Γ. See Chapter 3 for a fuller description of tMRCA

estimation and the mean rate prior for the relaxed clock analyses.

6.5 Results

6.5.1 Two distinct clusters of sequences in *env* and *gag* sequence sampling

Sampling of sequences at pre-treatment time points in P1 revealed additional complexity to the two clusters of sequences identified in the population sequences. In particular, recombinants of the original and co-circulating strains were identified, as well as minority variants of the original and co-circulating strains. Sequences sampled by bacterial cloning for *gag* and SGA sequences for *env*. The number of sequences sampled at each time point for each gene and technique are listed in Table 6.3.

Table 6.3 The number of partial- or full-fragment sequences sampled for each gene/technique at each time point. *Bracketed numbers are the number of clones obtained with each primer set.*

Sample (Days post-exposure)	P1 *gag* clones	P2 *gag* clones
63	29 (14/15)	27 (10/17)
137	6 (6)	8 (8)
185	18 (8/10)	…
235	40 (40)	11 (11)
397	…	…
439	28 (21/7)	…
474	17 (17)	…
819	44 (14/18/12)	…
Sample (Days post-exposure)	P1 *env* amplicons	P2 *env* amplicons
63	21	19
137	…	…
185	…	…
235	17	16
397	…	…
439	…	…
474	3	…
819	…	…

6.5.2 Phylogenetic analysis

6.5.2.1 Initial distance-based network evaluation of tree-like topology

If recombination had occurred following superinfection (Streeck et al. 2008), it would violate some assumptions of the models used for phylogenetic tree analysis. Therefore, I initially screened the unstripped *env* SGA and *gag* clonal sequences for tree-likeness. The NeighborNet method (Bryant and Moulton 2004) provided the best fit for the data (least squares fit 99.7%). NeighborNet analysis of the unstripped *env* fragment revealed that the topology was tree-like, with two P1 day 235 sequences located some distance from the P1 and P2 clusters, with bootstrap support of 100%. For the *gag* network, the 1983 subtype B reference sequence (HXB2) and a 1983 subtype D sequence were included as reference sequences for comparison, as the day 63 clonal samples for P1 and P2 were almost identical (Chapters 3 and 4). The network revealed a roughly tree-like structure with two clusters of sequences separated with bootstrap support of 98.2%. 'Cluster 2' did not contain any sequences from P2 or any sequences from P1 at days 63 or 137. Sequences from days 185 and 235 were located in both clusters.

6.5.2.2 Maximum likelihood and Bayesian MCMC based trees

I first constructed ML and Bayesian MCMC based trees for the long *env* fragment (Figure 6.5) using the GTR+Γ+I substitution model and reference sequences from SPARTAC and the LANL database (LANL, US). The topology of both trees for P1 and P2 was consistent with that seen in the network. Two P1 sequences from day 235 sat at some distance from the other P1 and P2 cluster with bootstrap and posterior probability supports of 100%. The bootstrap support for the day 235 sequences clustering with the main P1P2 cluster was

96% in the ML tree. In the Bayesian MCMC based tree, the bootstrap support was only 58.0%, probably due in part to the inclusion of fewer reference sequences in order to obtain a result in under 14 days (71 for Bayesian MCMC versus 147 for ML), however rerunning the ML analysis with the smaller alignment only reduced the bootstrap support for that node to 91.5% (data not shown). It was probable that the two most divergent sequences at day 235 had originated from an epidemiologically unlinked host but it was also possible that one or both day 235 sequences had recombined with the original infection.

I constructed ML and Bayesian MCMC based trees for the clonal *gag* fragment sequences (Figures 6.6 and 6.7 respectively). The sequences clustered together to the exclusion of all sequences with bootstrap support of 100%. The topology within the cluster agreed with the NeighborNet method, with the same Cluster 2 sequences branching from the other P1 and P2 sequences, with a bootstrap of 99.8%.

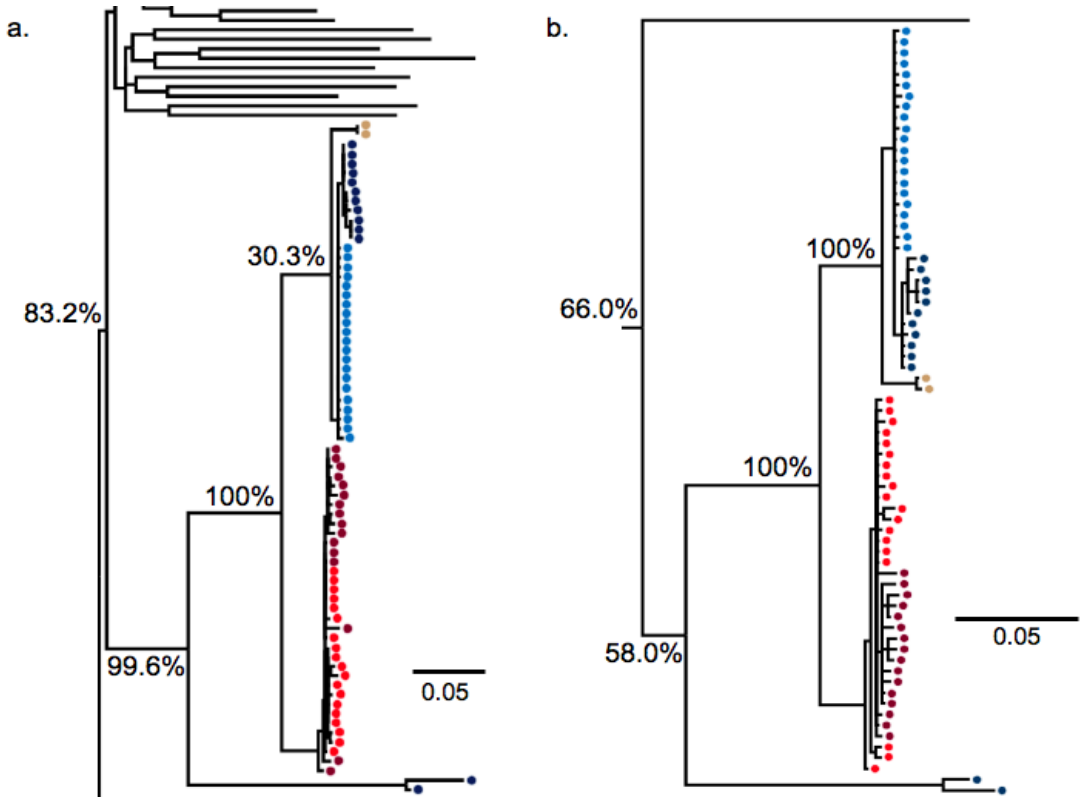


Figure 6.5 Zoomed-in views of a. ML and b. Bayesian MCMC based trees of the SGA long* env *fragment sequences at all sampled time points for P1 and P2. *P1 colours: intermediate blue=day 63, dark blue=day 235, light brown=day 474. P2 colours: light red=day 63, dark red=day 235. Bootstrap values for cluster support at internal nodes are given as percentages. The scale bar shows genetic distance, so that 0.05=5%. Fewer reference sequences were used to construct the Bayesian MCMC based tree (71 for Bayesian MCMC versus 147 for ML).*

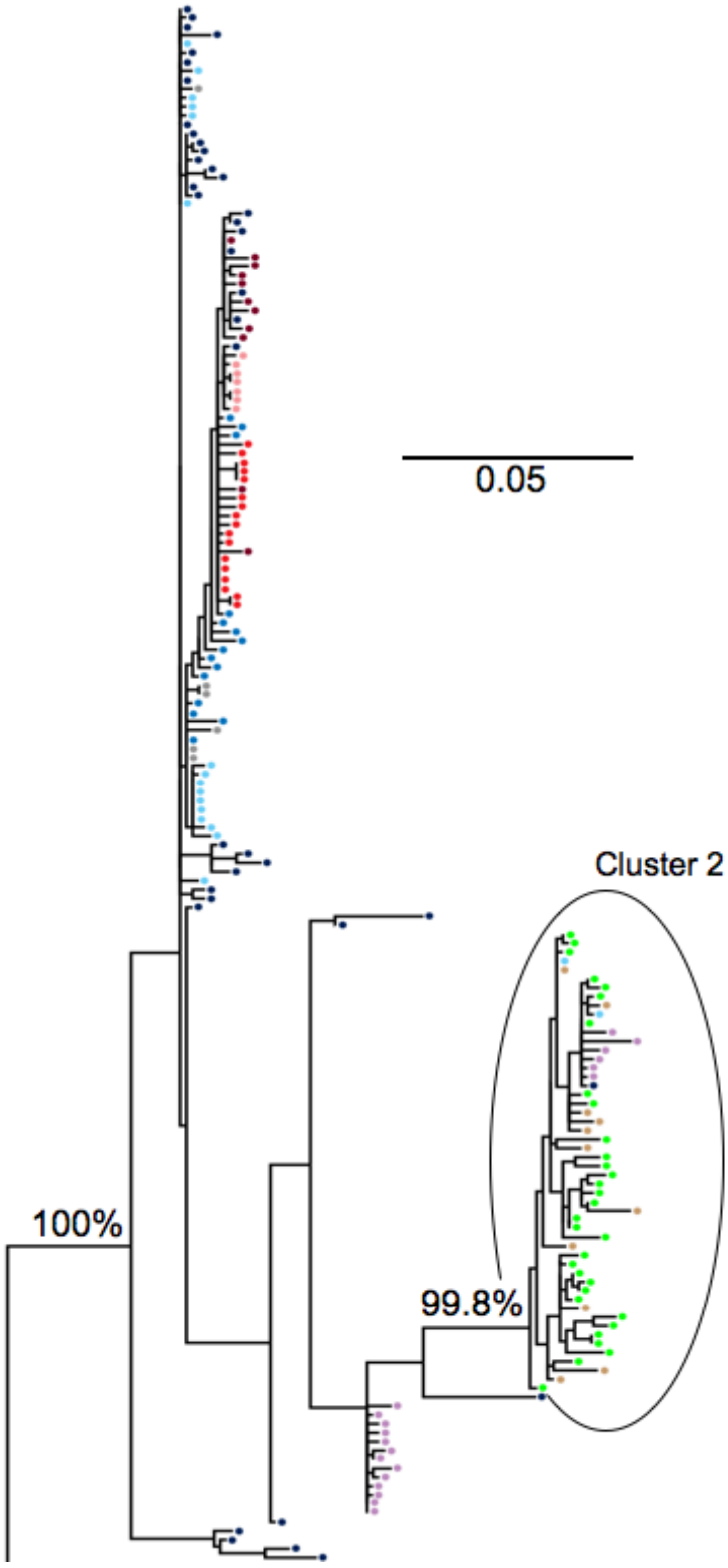

0.05
Cluster 2
100%
99.8%

Figure 6.6 (previous page) Zoomed-in ML tree of clonal* gag *sequences from P1 and P2. Coloured labels identify sequences from different patients and different time points. *For P1, the colours are: intermediate blue=day 63, grey=day 137, light blue=day 185, dark blue=day 235, purple=day 439, brown=474, green=day 819. For P2, the colours are: light red=day 63, pink=day 137, dark red=day 235. Bootstrap values for cluster support are given as percentages. The scale bar shows genetic distance (0.05 = 5%).*

As with *env*, the phylogenetic tree analysis was repeated using a Bayesian MCMC based approach but with a smaller (15 sequences) reference alignment than the ML tree (91 sequences) so that the analysis finished in under 14 days. Again, as with *env*, repeating the ML alignment with the smaller *gag* alignment did not affect the clustering of the P1 and P2 cluster, including the epidemiologically unlinked sequences (99.7% for the smaller alignment versus 100% for the larger alignment; data not shown).

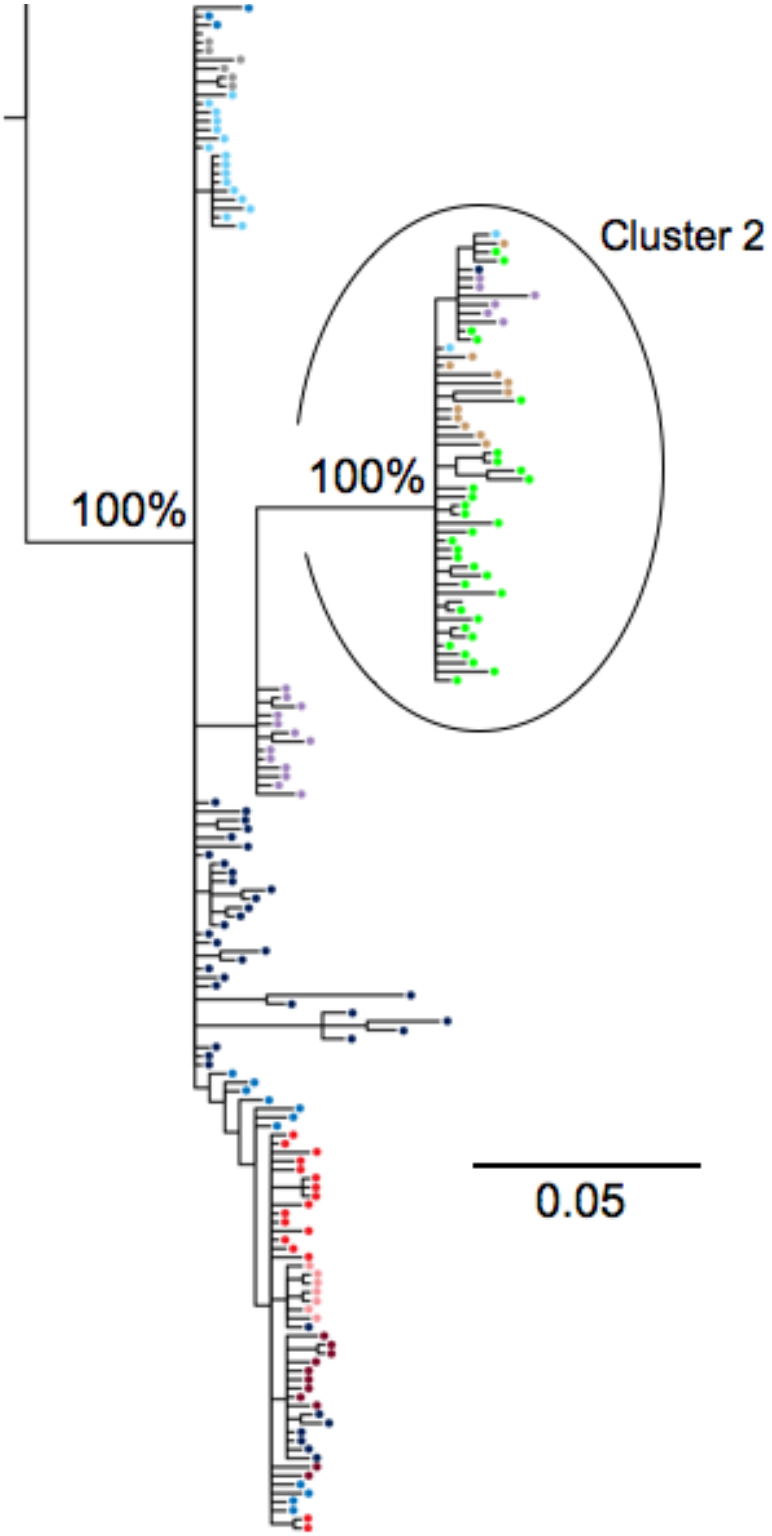

Cluster 2
100%
100%
0.05

Figure 6.7 (previous page) Zoomed-in Bayesian MCMC based tree of clonal* gag *sequences from P1 and P2. *Coloured labels identify sequences from different patients and different time points. For P1, the colours are: intermediate blue=day 63, grey=day 137, light blue=day 185, dark blue=day 235, purple=day 439, brown=474, green=day 819. For P2, the colours are: light red=day 63, pink=day 137, dark red=day 235. Bootstrap values for cluster support are given as percentages. The scale bar shows genetic distance (0.05 = 5%). Cluster 2 is indicated.*

6.5.3 Identification of recombinants in *env* and *gag*

Since I had used single genome amplification (SGA) for sampling of *env* sequences, detection of recombinant sequences probably reflected true recombination between the epidemiologically unlinked strains rather than being an artefact of PCR or consensus sequencing (Jordan et al. 2010; Salazar-Gonzalez et al. 2008). The recombinant sequence detected at day 235 (day 235 P1 SGA sequence 13; Figure 6.8) and an example of one of the recombinant sequences detected at day 474, using the most divergent unstripped sequence at this time point (day 474 P1 SGA sequence; Figure 6.9) are shown in the output from SimPlot v3.5.1 (Lole et al. 1999), relative to the day 63 consensus sequence and the most divergent sequence, detected at day 235. For *gag*, recombinants were also detected but the recombinants may have been an artefact of the bacterial cloning method in the presence of multiple infection (data not shown).

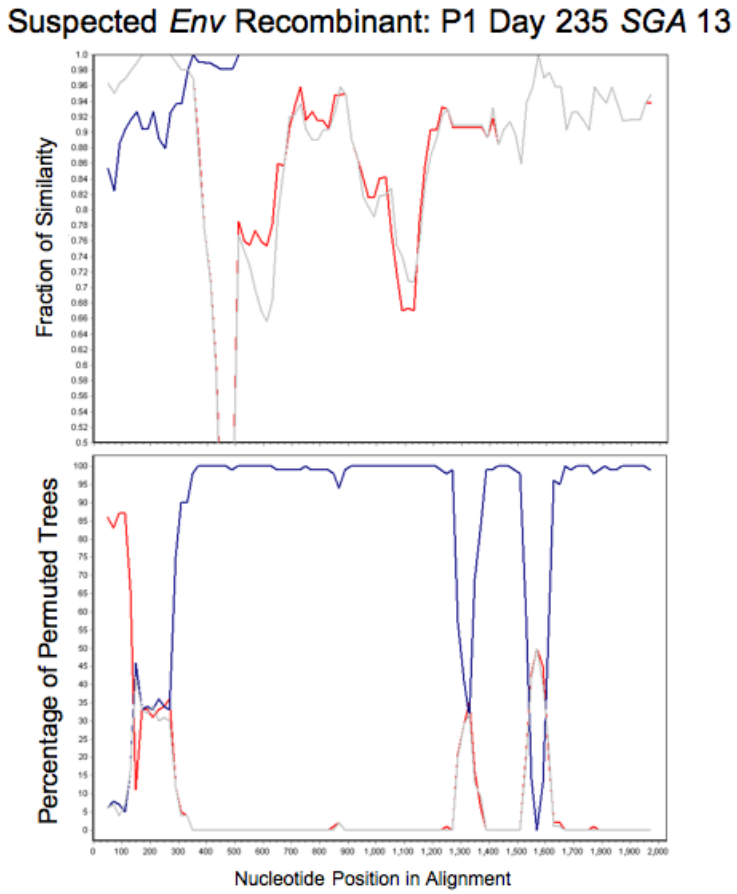


Figure 6.8 Query-versus-reference plots for the suspected recombinant sequence P1 Day 235* env *SGA 13. *Plots show (top) the fraction of sequence similarity (out of 1.0) and (bottom) the percentage of permuted trees that support clustering with the reference sequence. The red line represents the reference consensus sequence for P1 at day 63. The blue line represents the reference sequence P1 day 235* env SGA *14. The purple line represents the reference sequence P1 Day 439* gag *01. The grey line represents the reference consensus sequence for P2 at day 63.*

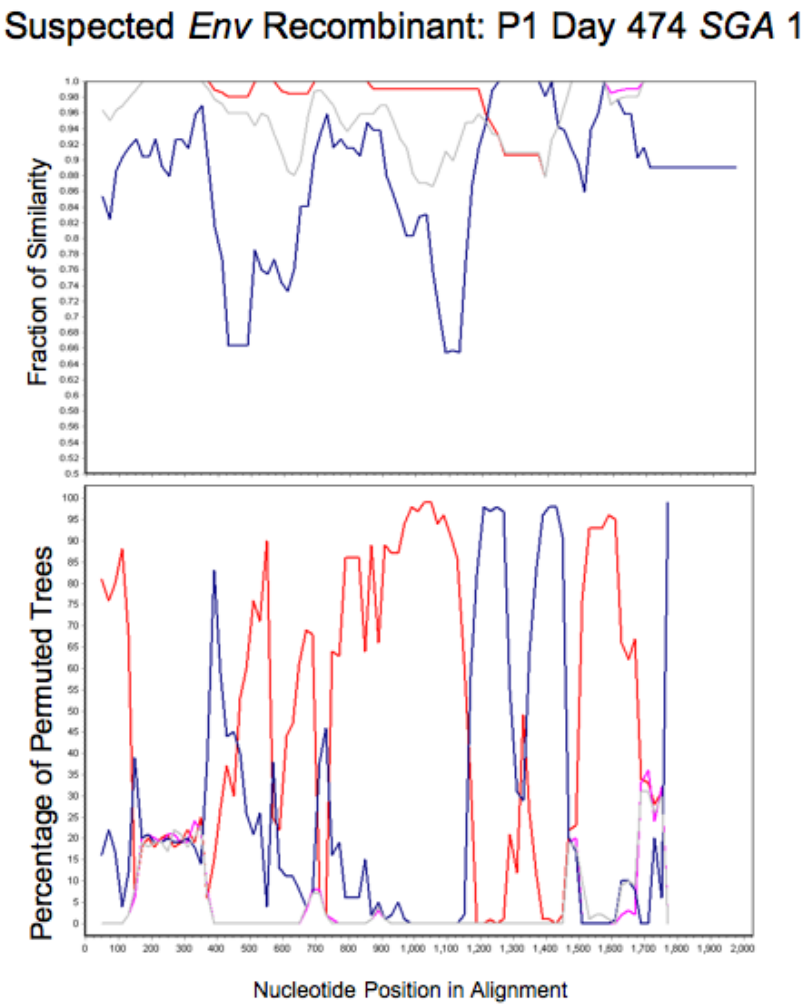


Figure 6.9 Query-versus-reference plots for the suspected recombinant sequence P1 Day 474* env *SGA 1. *Plots show (top) the fraction of sequence similarity (out of 1.0) and (bottom) the percentage of permuted trees that support clustering with the reference sequence. The red line represents the reference consensus sequence for P1 at day 63. The blue line represents the reference sequence P1 day 235* env SGA *14. The purple line represents the reference sequence P1 Day 439* gag *01. The grey line represents the reference consensus sequence for P2 at day 63. The purple line represents the reference consensus sequence for P1 at day 235.*

6.5.4 Quantification of within-host *env* diversity in terms of the tMRCA

Within-host diversity was quantified using a Bayesian MCMC based approach to demonstrate that viral diversification in P1 after day 63 was consistent with invasion by an epidemiologically unlinked variant. Recombination had been previously shown to bias

coalescence analyses by over-estimating the tMRCA, as recombination increases the amount of apparent evolution if a substitution-only model is used (Lemey et al. 2004; Lemey et al. 2009). However, I merely wanted to test the null model that the amount of diversification was due to within-host evolution P1, so I calibrated the substitution rate to the amount of evolution observed. Relaxed molecular clock analyses failed to converge, even when enforcing a strong prior on the rate (Chapter 4) suggesting that recombination might be causing apparent rate heterogeneity among branches. Strict molecular clock analyses did converge at an estimate of divergence at least 8.5 years prior to exposure to the original donor. Although the precise reliability of the strict clock estimate was questionable, it was probable that the estimate was consistent with co-circulation of epidemiologically unlinked strains.

6.5.5 Immunological analysis

6.5.5.1 Class I HLA-mismatched mutation analysis in the gag protein sequence

As in Chapter 5, I initially looked for mutations in the gag sequences from P1 and P2 that might indicate transmission of Class I HLA-mismatched mutations from the donors that may affect replication capacity (Brockman et al. 2010; Miura et al. 2009a). P1 possessed HLA-A*0201, A*2402, HLA-B*1402, B*3543, Cw*0102, Cw*0802; P2 possessed HLA-A*0101, A*2901, B*0801, B*5001, Cw*0602, Cw*0701. Neither participant possessed HLA types that are strongly associated with protection from progression in chronic infection (X. Gao et al. 2001; Kiepiela et al. 2004). Again, I failed to find any protein mutations in the clonal sequences from P1 or P2 that were known to be associated with either elite control of viral load *in vivo* or reduced replication rates *in vitro* (Brockman et al. 2010; Miura et al. 2009a; Miura et al. 2009b). However, I confirmed that P1 had acquired a mismatched HLA-associated mutation in the suspected invading sequence within the epitope TSTLQEQIGW (TW10). The TW10 mutation TSTLQEQIAW (G248A) had been associated with hosts carrying the HLA-B*5801 allele (Martinez-Picado et al. 2006), but neither P1 nor P2 carried this allele. Clonal sequences confirmed that both P1 and P2 were infected with the wild-type TW10 mutation before week 235.

6.5.5.2 Acquisition of a Class I HLA-matched escape mutation by P1

As in Chapter 5, I then looked at the SLYNTVATL (SL9) epitope in the gag protein, which was known to be restricted by HLA-A*0201, and because P1 carried the HLA-A*0201 allele while P2 did not. Specific SL9 mutations had also been associated with lower viral loads *in vivo* during chronic infection (Iversen et al. 2006; Karlsson et al. 2007). At day 63 and day 137, all sequences from both P1 and P2 contained the wild-type SL9 epitope. A minority of variants at days 187 and 235 in P1 contained the known SLFNTVATL (Y79F) mutation for escape from HLA-A*0201 (Goulder et al. 1997; Harrer et al. 1998; Iversen et al. 2006; Jamieson et al. 2003) instead of the wild-type SL9 epitope, and these sequences had been inferred to be recombinant sequences (data not shown).

6.5.5.3 CTL response to SL9 or variants

Despite carrying the HLA-A*0201 allele, P1 did not mount a detectable CTL response on ELISpot to either the wild-type SL9 or previously described Y79F-containing mutants at days 63, 235 or 474 (Table 5.2). Although generation of the escape mutation within P1 could not be excluded, the appearance of the Y79F mutation would be consistent with invasion by an epidemiologically unlinked infection already carrying the escape mutation at SL9 (Chapters 5 and 8).

6.5.5.4 Humoral immune response

The most likely source of selection driving the evolutionary rate in *env gp120* would be evasion of neutralizing antibodies. No neutralizing antibodies were detected against *env* at day 185, which is when the first Cluster 2 *gag* sequences were first detected. Please see Chapter 3 for a full description of this result.

6.5.5.5 Potential N-linked glycosylation sites in gp120

No neutralization data were available for serum from 6 months and 1 year. The day 474 sequences shared 15 of the 17 PNLGSs present in both P1 and P2 at day 63, compared with 7 for greater than 90% of subtype B SPARTAC sequences. The two most divergent sequences at day 235 shared 13 of the 17 PNGLSs present in both P1 and P2 at day 63, using a conservative AA alignment. The two most-divergent env SGA sequences SGA from day 235 and the two SGA sequences from 474 shared inferred regions of recombination in the V4 region, coinciding with the inferred region of recombination between these two strains. The mean number and range of PNLGSs across V1V4 For each participant and time point is shown in Table 6.4.

***Table 6.4 Number of PNLGSs at different time points in the* V1V4 *region of* env gp120**

	PNLGSs by Participant	
	P1	P2
Day 63 median	19 (18 to 20)	24 (23 to 25)
Day 235 median	19 (18 to 20)	24 (22 to 25)
- Divergent day 235 sequences	18 and 18	…
Day 474 sequences 1 and 2	18 and 19	…

Having inferred that the divergent day 235 sequences (sequences 13 and 14) shared a similar *V4* region with the day 474 sequences, I examined the inferred amino acid sequence for the C3-V4 region in more detail (Tables 6.5 and 6.6).

Table 6.5 Aligned AA sequences, from the beginning of the C3 region until HXB2 env AA position 376. Sequence name abbreviations are: $P1_{con.}$=consensus sequence for P1, $P2_{con.}$=consensus sequence for P2, $P1_{SGA}$=SGA sequence from P1. Mutations are highlighted in red and underlined, relative to the P1 consensus sequence from day 63, which is highlighted in black bold. The names of the two most divergent SGA sequences in P1 at

day 235 are highlighted in blue. The HXB2 subtype B reference sequence is given for comparison.

Name	Partial C3 AA sequence (HXB2 env AA 331 to 376)
HXB2	CNISRAKWNNTLKQI-ASKLREQFGNNKTIIFKQSSGGDPEIVTHSFN
Day 63	
-P2$_{con}$	CNVNGTKWNQII-QKVAKKLRAQF-NNKTIVFNHSSGGDLEIVMHSFN
-P1$_{con}$*	**CNVSGN-WSQIL-QEVAKKLSAQF-ENKTIVFNHSSGGDPEIVMHSFN**
-P1$_{SGA}$21	CNVSGN-WSQIL-QEVAKKLSAQF-ENKTIVFNHSSGGDPEIIMHSFN
Day 235	
-P2$_{con}$	CNVNGTKWNQIL-QKVAKKLRAQF-NNKTIVFNHSSGGDLEIVMHSFN
-P1$_{con}$	CNVSGN-WSQIL-QEVAKKLSAQF-ENKTIVFNHSSGGDPEIVMHSFN
-P1$_{SGA}$1	CNVSGN-WSQIL-REVAKKLSAQF-ENKTIVFNHSSGGDPEIVMHSFN
-P1$_{SGA}$3	CNVSGN-WSQIL-QKVAKKLSAQF-ENKTIVFNHSSGGDPEIIMHSFN
-P1$_{SGA}$4	CNVSGN-WSQIL-QKVAKKLSAQF-ENKTIVFNHSSGGDPEIIMHSFN
-P1$_{SGA}$5	CNVSGN-WSQIL-QKVAKKLSAQF-ENKTIVFNHSSGGDPEIIMHSFN
-P1$_{SGA}$13	CEVNETAWNKNLKEI-AGKLRKQFG-NKTIAFNSSGGDPEIVMHSFN
-P1$_{SGA}$14	CEVNETAWNKNLKEI-AGKLRKQFG-NKTIAFNSSGGDPEIVMHSFN
Day 474	
-P1$_{SGA}$1	CNVSGN-WSQIL-QKVAKKLSAQF-ENKTIVFNHSSGGDPEIVMHSFN
-P1$_{SGA}$2	CNVSGN-WSQIL-QKVAKKLSAQF-ENKTIVFNHSSGGDPEIVMHSFN

The C3V4 protein region is important for CD4 binding (Moore et al. 2009), and the C3 region of HIV-1 had previously been shown to be highly-exposed but to be also functionally important for viral infectivity (C. Otto et al. 2003). In Chapter 3, virus isolated from P1 was shown to have a higher infectivity than the virus in P2. It was observed that the C3 region of the day 474 AA sequences in P1 was

Table 6.6 Aligned sequences for the env V4 region (HXB2 env AA 385 to 418). Sequence name abbreviations are: P1$_{con.}$=consensus sequence for P1, P2$_{con.}$=consensus sequence for P2, P1$_{SGA}$=SGA sequence from P1. Mutations are highlighted in red and underlined, relative to the P1 consensus sequence from day 63, which is highlighted in black bold. The names of the two most divergent SGA sequences in P1 at day 235 are highlighted in blue. The HXB2 subtype B reference sequence is given for comparison.

Name	Full V4 AA sequence (HXB2 env AA 385 to 418)
HXB2	CNSTQLFNSTW-FN-S--TWS-TEGSNN-TEG-----S-D-TITLPC
Day 63	
-con. P2	CNTTQLFNSTW--NDTN—TWNDTNIGNNNT-G-----N--ITLPC
-con. P1*	**CETTQLFNSTW--N----T---TQLFNNDT—G-----SED--ITLPC**
Day 235	
-con. P2	CETTQLFNSTW--N----T---TQLSNNDT-G-----N--ITLPC
-con. P1	CETTQLFNSTW--N----T---TQLSNNDT-G-----SED--ITLPC
-$P1_{SGA}1$	CETTQLFNSTW--N----T---TQLVNNDT-G-----SED--ITLPC
-$P1_{SGA}3$	CETTQLFNSTW--N----T---TQLFNNDT-G-----SED--ITLPC
-$P1_{SGA}4$	CETTQLFNSTW--N----T---TQLFNNDT-G-----SED--ITLPC
-$P1_{SGA}5$	CETTQLFNSTW--N----T---TQLFNNDT-G-----SED--ITLPC
-$P1_{SGA}7$	CETTQLFNSTW--N----T---TQLANNDT-G-----SED--ITLPC
-$P1_{SGA}13$	CDTTPLFNSTWRVN-S--TWNETQ------G-----AENKTIILQC
-$P1_{SGA}14$	CDTTPLFNSTWRVN-S--TWNETQ------G-----AENKTIILQC
Day 474	
-$P1_{SGA}1$	CDTTPLFNSTWGVN-S--TWNETQ------GYFPELAE---I-LQC
-$P1_{SGA}2$	CDTTPLFNSTWGVN-S--TWNETQ------G-----AENKTIILQC

almost identical to that in all of the day 63 AA sequences in P1, except for an E to K mutation at HXB2 position 345. By contrast, both sequences sampled at day 474 had V4 regions that were more similar to the most divergent sequences at day 235. The C3 peptide to which P1 had a CTL response at day 63 only was conserved in all of the SGA sequences (Chapter 5; data not shown).

6.6 Discussion

6.6.1 Recombination demonstrated by SGA in *env*

6.6.1.1 Recombination occurs in env *during multiple infection*

The *env* nucleotide sequence data was collected using SGA, so any recombinants detected are likely to be true recombinants (Jordan et al. 2010; Salazar-Gonzalez et al. 2008). At day 235, a minor variant was detected in P1 that was quite different to the original infecting strain in either P1 or P2. That the minor variant likely represented an epidemiologically unlinked strain, rather than within-host diversification, was consistent with the tMRCA estimated by molecular clock analysis. There was also evidence of recombination between the invading strain and the strain of the original infection at days 235 and 474, although the pattern of recombination breakpoints at days 235 and 474 were quite different. Inspection of the V4 region amino acid sequences demonstrated persistence of non-synonymous changes at day 474 from the minor variant at day 235, although each day 474 variant was slightly different, suggesting that the recombination breakpoint pattern represented phenotype selection and diversification in V4.

6.6.1.2 The detected strains at day 474 more closely reflect the strain of the original infection rather than the invading strain

Both recombinant strains at day 474 closely resembled the original strain except in the *V4* region. Since the V4 region was largely cut from the alignment used to reconstruct the phylogenetic tree in Figure 6.2 due to the number of insertions and deletions, the day 474 clustered more closely with the original strain than the two outlier sequences at day 235. Without more SGA sequences from

day 474 or SGA data from day 819, it is difficult to know if the two recombinants detected at day 474 represent a stable superinfection by multiple strains with slight variation in the V4 region. However, the population-amplification sequence in *env* (Chapter 5) at day 819 were more closely related to the original strain in P1 than to the epidemiologically linked, original strain in P2 so one might expect that SGA sequences at day 819 would also most closely represent the original strain in P1. If one assumes the two detected sequences at day 474 represent the dominant strains at that time point, then there is sequence dominance by the original strain in *env*.

6.6.1.3 Potential functional and immunological explanations for recombination centring on V4

This conclusion suggests that there may be some quality of the invading V4 sequence that is relatively more beneficial to the virus than the sequence of the strain causing the original infection. *V4* is typically a very diverse hypervariable region in *env gp120* that is often particularly variable in its glycosylation patterns (Castro et al. 2008). Recombination may be consistent with escape from neutralizing antibody (Castro et al. 2008; Koelsch et al. 2003; Moore et al. 2008). There was no evidence of any significant, autologous neutralization activity towards the original infection by sera sampled at day 185, which was earliest time-point when epidemiologically unlinked sequences were detected in gag in Chapter 5. If one then assumes that the epidemiologically unlinked env sequence enters participant P1 before day 185, then there would be no detectable immune-evasion benefit, at least initially, to acquisition of the V4 region of the invading strain by the original strain.

6.6.1.4 Comparison with results seen in gag

What is clear is that the results in *env* contrast with those seen in *gag*, where the majority of sequences sampled after day 235 resemble the inferred invading strain (see 6.6.2). The different patterns seen between genes is consistent with the view that genetic subunits of the virus may act as agents of selection, rather than whole virus genetic haplotypes (Gardner and Grafen 2009). One would predict that multiple-infection would have different consequences for evolution in *env* and *gag*, consistent with the idea of genetic regions acting as units of selection rather than as whole viruses (Gardner and Grafen 2009). However, an important caveat to this conclusion is that the SGA approach to sampling sequence diversity in *env* used fewer primer sets in general than the clonal sequence sampling approach used in gag. Therefore, while the SGA approach in env is specific for detection of true recombination, it may be less sensitive for sampling the breadth of diversity, particularly if the primers were to bind less well to the invading strain or recombinants and contribute to the lower SGA yield at day 474.

6.6.2 Superinfection by a dominant strain demonstrated in *gag*

6.6.2.1 Differentiating between in vivo *and* in vitro *recombination in* gag

That *gag* data were collected by bacterial cloning using multiple primer sets in order to try and capture the full genetic diversity present at each time point because bacterial cloning allows for relatively efficient sampling with multiple primers compared with SGA. The main disadvantage of the bacterial cloning approach to sequence sampling is that the haplotype in the sequences may reflect artificially induced recombination created by the PCR process (Jordan et al. 2010; Salazar-Gonzalez et al. 2008). However, recombination detected in the population sequences (Chapter 5) and clonal sequences (Chapter 6) is more likely to be true *in vivo* recombination if the break points are reproducible across multiple primer sets and time points.

6.6.2.2 The superinfecting gag variant at later time points detected by population sequencing was confirmed by clonal sampling

The dominant variant cluster at day 819 in *gag* was detectable as a minor variant as early as day 185 in participant P1. This observation was consistent with the timing of HCV positivity in the same participant. However, unlike in *env*, there was no clear outlier cluster in gag from which to track the invasion of sequence or part-sequence, since the superinfecting strain appeared to be phylogenetically highly related to the original strain, although the within-host diversity in P1 was substantially greater from day 185 than even the maximum between host divergence between the day 63 P1 cluster and the whole of the P2 cluster up to day 235. Had data from P2 not been available, then multiple infection may not

have been detectable based upon the sampled *gag* sequences from P1 alone.

6.6.2.3 Immunological data in gag

I investigated whether apparent superinfection by a strain resembling the inferred epidemiologically unlinked strain had any potential replication or immune escape consequences for *gag*. I confirmed that the HLA-matched SL9 mutation (Y79F) appeared to be a feature of the invading strain because I did not find any sequences with this mutation that clustered with the original strain in P1 and P2. This supported my view (Chapter 5) that the mutation had been acquired from the epidemiologically unlinked donor rather than having arisen in P1. Furthermore, P1 did not made a detectable CTL immune response to this epitope over the first 474 days since exposure to the original host (Chapters 5 and 8). The fact that the SL9 mutant wins out may reflect genetic hitch-hiking (J. M. Smith and Haigh 1974) but it may also imply that the ELISpot is not sensitive to the true complexity of the host selection in *gag* suggested by studies of elite virus control (Blankson et al. 2007; Euler et al. 2012; Miura et al. 2008; Simek et al. 2009).

6.6.3 The qualitative diploid model allows predictions of HIV-1 evolution

6.6.3.1 Predictions that do no involve immune blocking of multiple infection

As the immunological data for both *env* and *gag* could not conclusively demonstrate an immunological advantage that would allow the invading strain, or at least genetic subunits of it, to invade the population, I turned to my simple qualitative model. Under the model, multiple-infection with eventual superinfection would occur if either the invading strain was substantially more fit than the original strain or if there was some advantage to having a recombinant mixed state. Since the majority of sampled sequences in both *env* and *gag* resembled the original infecting strain after day 185, I inferred that the scenario described by Figure 6.3 was the most plausible. In this scenario, the mixed recombinant state *AB* has, on average, the highest relative fitness and the relative fitness of pure original strain *AA* is greater than that of pure invading strain *BB*. Thus, once a steady state is reached, the viral population will on average resemble phylogeny *A* more than phylogeny *B*. Of course, it is unlikely that a steady state is ever truly reached due to constant changes in the host-pathogen interaction. However, it is quite possible that the dominant, probable-recombinant strains detected at later time points in both *env* and *gag* have triumphed due to being the fittest recombinant out of a swarm of recombinants that were on average fitter than both the original and invading strains. (S. P. Otto and Day 2007)

6.6.3.2 The role of the mixed recombinant state

If “fitness” cannot be solely attributed to an immune response that

favours either the original of invading strain sequences, or to some other quality such as replication capacity, then it may be that the mixed nature of the recombinant state itself confers benefit to the virus, by facilitating rapid adaptation to host conditions. The benefits of the mixed recombinant state may not be limited to invasion in multiple-infection, but may also be of relevance during primary infection by the initial strain. If diploidy and the recombinant state are key to HIV-1 virus population evolution and adaptation within hosts, then blocking the ability of the virus to recombine may be a novel way of arresting early infection.

6.7 Conclusion

6.7.1 Future work

My data and the model suggest at least two avenues for further response. Firstly, further longitudinal characterization of *env* recombination, infectivity and the host neutralizing antibody response immune response may shed light upon the average fitness of the mixed recombinant state in *env* (Bailey et al. 2006; Euler et al. 2012; D. L. Kothe et al. 2007; Mascola et al. 2000; Wei et al. 2003; Weiss et al. 1985; Yeh et al. 2010). Secondly, characterization of the mixed recombinant state in *gag*, particularly its average replication capacity *in vitro* may elucidate the contribution of mechanisms other than the immune response to facilitating or preventing multiple infection (Brockman et al. 2007; Miura et al. 2008; Miura et al. 2009a; Miura et al. 2009b).

6.7.2 Implications of my data for transmission and primary infection

If *gag* is prone to superinfection whereas the detection of *env* diversity is a more stochastic process dependent upon the degree of transmitted *env* diversity and it ability to undergo homologous recombination, then this may also partly explain why initial infections are usually homogeneous after sexual transmission. Therefore, apparently homogeneous infections may not be initiated by a single strain and further work is required to understand the complexity of within-host HIV-1 evolution.

Chapter 7 Discussion

7.1 Overview

In this thesis, I have explored the complex stochastic and adaptive features of HIV-1 transmission. In Chapters 3 to 6, I have not only presented a rare incidence of contemporaneous transmission of divergent viruses from a single donor to two recipients, but also the subsequent acquisition of an epidemiologically unlinked strain by one of the recipients, resulting in superinfection. I have also demonstrated the presence of recombinant strains following multiple-infection. I have discussed the immunological implications of divergent variant transmission as well as the acquisition of potential escape mutations through multiple-infection. I have also addressed the analytical challenges of investigating complex transmission scenarios within the same subtype. In Chapter 7, I will revisit the two questions that I posed at the conclusion of Chapter 1 in light of my findings.

7.2 Key questions revisited

7.2.1 Is there any evidence of hard selection of viral diversity at transmission?

7.2.1.1 Brief discussion of context for this study

Great efforts have been made to characterize transmissible viral variants in the search for an efficacious, preventative HIV-1 vaccine strategy (Pope and Haase 2003). Before I commenced the work presented in this thesis, previous authors had suggested that the genetic bottleneck seen at transmission resulted from hard selection at the mucosal barrier, resulting in heterogeneous viral populations after most cases of sexual transmission (Keele et al. 2008; Salazar-Gonzalez et al. 2008). Furthermore, it had been suggested that transmitted variants might be uniquely infectious, and so great efforts had and continue to be made to identify and characterize the founder virus or viruses (Keele et al. 2008; Pope and Haase 2003; Salazar-Gonzalez et al. 2008). However, many of the assumptions underpinning the associated models of transmission in these studies contained assumptions about viral behaviour that could not be directly verified for ethical or technical reasons.

7.2.1.2 Rare case study presented in this thesis

Therefore, I undertook to see if these assumptions would hold for a special case of transmission. In Chapter 3, I presented a rare case of inferred, contemporaneous transmission from a single donor to two recipients. However, it was clear that only 63 days after the single episode of common exposure the participants were infected with very different viral variants it at least two genes, *env gp120* and *nef*. In Chapter 4 I inferred that the degree of divergence in each strain could be consistent with a single, chronically infected donor or the

transmission of epidemiologically linked strains from a single donor. Furthermore, it was highly unlinked that the divergence had developed within the recipients, but more probably because of diversifying selection within the donor or the donor's donor in these regions (Herbeck et al. 2011).

7.2.1.3 Between-host divergence

The most divergent genomic region between the two recipients was *env gp120*, and I inferred that the sequence differences could potential lead to significant differences in antigenicity through alterations in the hypervariable regions and the glycan shield, while remaining consistent with epidemiological linkage through a common donor. However, no evidence of an autologous or heterologous neutralizing antibody responses to the day 63 post-exposure env variants was detected in either participant as late as day 185 post-exposure. Furthermore, differences in infectivity related to mutational differences in env did not explain how each participant came to be infected with different envelope strains. In addition, I could not attribute the differences between the hosts in the most conserved gene, *gag*, to detected CTL responses.

7.2.1.4 Lack of evidence of hard selection due to absence of an immune response

Neutralizing antibody responses in the peripheral blood and CTL responses have been used as surrogate markers of a protective immune response against infection (Burton et al. 2004; Letvin 2006; Y. Liu et al. 2011; Saphire et al. 2001). While non-neutralizing antibodies may play some role in controlling viraemia, neutralizing antibodies are the most potent at inhibiting cell-to-cell viral transmission (Albert et al. 1990; Walker et al. 2009; Wei et al. 2003).

We did not look at mucosal-specific immunity (Mazzoli et al. 1997; Tudor et al. 2009) because this compartment had not been sampled, but we reasoned that a failure to elicit neutralizing antibodies in the blood probably reflected the low immunogenicity of the *env* variants generally.

7.2.1.5 Alternative causes for selection of viral variants

Another possibility is that hard selection may be unrelated to antibodies but may instead reflect subtle differences in the different abilities of certain variants to enter cells. However, P1 and P2 had markedly different envelopes with different glycosylation profiles that demonstrated different infectivity levels in vitro. Although our study was not set up to look specifically at infectivity, the different genetic and phenotypic differences in the two variants demonstrated that ate least twp variants were capable of successful transmission within the donor. Furthermore, the degree of difference was greater than that seen in co-infected participants sampled by previous authors.

7.2.1.6 Conclusion and a further question

My findings suggested that a single donor can transmit very different variants contemporaneously despite and that hard selection due to an immune response in the recipients does not account for this difference. How does one then explain the extreme bottleneck seen in most early infections? Further research is needed to answer this question.

7.2.2 To what extent does recombination play a role in shaping viral diversity when two variants of the same subtype co-circulate at very different initial frequencies?

7.2.2.1 Co-circulation of an initial infection and an invading strain

within the same host for several months

I demonstrated that two very different variants as well as recombinants may co-circulate within the same host for several months. Therefore, it is conceivable that more highly related strains may co-circulate in the acute phase but at very different frequencies within the population.

7.2.2.2 Consideration of the role of recombination in viral adaptation to a new host and its implications for studying the invasion of a new host by a viral strain

A key assumption of the single variant transmission theory is that all within-host diversity seen during or shortly after the exponential phase of viral growth can be attributed to mutations away from a single founder strain in the case of homogeneous infections. However, I observed that the appearance of multiple infection in P1 was associated with the appearance of HLA matched immune escape mutations but that the sequences still clustered very close together on the phylogenetic tree. If had not been for the presence of the sequences in P2, it may have appeared as though P1 were evolving quickly, rather than that invasion of an exogenous strain and recombination had taken place. Recombination is more likely between highly related strains on a mechanistic level, and the degree of relatedness is known to affect the rate of observed recombination in *gag*. Likewise, homologous recombination is predicted to be low in genes that have a high rate of mutation by insertion or deletion, such as *env*. In the case of P1, recombination around the hypervariable regions facilitated the emergence of a strain incorporating the invading-strain's hypervariable region, in a pattern reminiscent of antigen-shifts seen in influenza.

7.2.2.3 Difficulties in studying recombination also suggest why it is useful for the virus in terms of Darwinian selection

Recombination is difficult to study within hosts with any statistical certainty, as it is more likely to occur if the strains are more highly related, but an increased rate of recombination between highly related sequences would allow for the sharing of beneficial mutations with closer 'kin'. Sharing benefits with closely related strains is consistent with predictions from inclusive fitness theory that organisms being acted upon by natural selection distribute benefits to a greater or less extent depending upon the degree of genetic relatedness between interacting individuals. In other words, a strain may try to directly and indirectly affect the success of reproduction of others in the population so as to optimize its own reproductive success. The effect that a strain has on other strains is modified by its degree of relatedness to other strains. However, relatedness does not necessarily involve kinship by common descent, but may involve sequences that become more similar through recombination.

7.2.2.4. The contribution of diploid models to studying within-host HIV-1 evolution

Diploid models may provide key qualitative insights into the evolution of a virus population when two strains co-circulate at very different initial frequencies. In particular, the role of the mixed recombinant state in facilitating the invasion of an epidemiologically unlinked strain. Since work was undertaken on this thesis, other authors have speculated about the role of certain viral mutations, genes and immunogenic regions to undermine the immune response to more conserved regions of the genome, in a process that some

authors call deceptive imprinting (Nara and Garrity 1998; Nara et al. 2010; Tobin et al. 2008) and others call the use of immunodominant decoys (Rolland et al. 2007; Rolland et al. 2010). It is plausible that multiple-infection broadens the immune target, such that the developing immune response becomes distracted from conserved regions. It is also plausible that the dominant virus during acute infection serves as a decoy so that the immune system begins to target the dominant strain just as the population is collapsing from a fall in the carrying capacity. The slower growing variants may then form a pool of diversity that the virus may use to escape from the immune response.

7.2.2.5. Extrapolating observations from multiple-infection to making predictions about viral evolution during acute infection

If very different variants may co-circulate for some time in hosts with multiple-infection, then it is also probable that co-circulation may occur in acutely infected hosts during acute infection but with much greater differences in genetic frequency. However, the apparent homogeneity of the 'env' of the infection may relate to the initial dominance of a particular strain and the limited ability of env variants to recombine to high frequency if mutations by insertion or deletion limit homology. Although I inferred that P1 and P2 were infected by the same donor, there was considerable heterogeneity on the hypervariable region of the *env* genes. However, in a gene such as *nef*, which is less prone to mutations by insertion or deletion, one may expect to see greater heterogeneity unless functionality of the nef protein puts strong selection on one variant over another. When I looked at multiple-infection in P1, I noted that envelope did not rapidly recombine. Rather the dominant variant at day 235 was the original strain and this may also have been the case at one and two years.

7.2.2.6 How my observations and model question the assumptions of previous models

Therefore, unlike in previous studies (E. E. Giorgi et al. 2010; Keele et al. 2008; H. Y. Lee et al. 2009; Ochsenbauer et al. 2012), I would not completely rely on *env* neutrality to provide the signal of co-infection by more than one strain during the acute phase, if insertion or deletion mutations inhibit recombination. In addition, aside from *nef*, in both *env* and *gag* the supposedly epidemiologically unlinked strains clustered together with the

original strains and the recombination signal was not definite in the bootstrap analysis. Thus the initially dominant strain might be dragging other strains closer to it in terms of similarity through recombination at more conserved motifs within the subtype. In acute infection, the initial degree of similarity might influence the ability of the sequences to come closer together and give the impression of a homogeneous infection because of the massive outgrowth of the dominant strain.

7.3 Further work and future directions

7.3.1 Replication capacity comparison in multiple infection using the gag-protease region

In Chapter 5, I discussed the utility of measuring relative replication capacity to assess whether the superinfecting strain was indeed fitter than the original infection. A recent study had looked at replication capacity in the setting of superinfection using only the pol gene, because of the possible role of transmitted drug mutations in altering viral fitness. However, the logical next-step is to investigate whether differences in the replication capacity attributable to the *gag-pol* region lead to superinfection, given that this is the most conserved region in the genome and targeted by the CTLs in P1 at least as early as day 63.

7.3.2 Confirming the role of the mixed recombinant state in facilitating multiple-infection

If there is no difference in the replication capacity between the original and superinfecting strain, or if the invading strain proves to replicate less readily *in vivo*, then the reason why the strain is able to take hold might be a combination of a fitter mixed recombinant state over many months, producing sustained apparent multiple-infection, as well as the eventual selection of one variant over the other by the host immune system. Therefore, it would be of grate interest to explore the significance of the mixed state for viral fitness and confirm predictions made by a simple diploid model. This may be of particular relevance to acute infection, if transmission of greater diversity, initially masked and shaped by the dominant strain, is in fact a viral strategy for adapting to the host.

7.4 Concluding remarks

It is probable that researchers have barely scratched the surface of the complex behaviours that a small-genome retrovirus can demonstrate, in order to adapt successfully for transmission between human hosts. The problem of HIV-1 transmission between hosts should be investigated and interpreted iteratively from experimental and theoretical angles in order to understand the exquisite adaptation of this pathogen for transmission.

Dark Helmet: Careful you idiot! I said across her nose, not up it!

Gunner: Sorry sir! I'm doing my best!

Dark Helmet: Who made that man a gunner?

Major Asshole: I did sir. He's my cousin.

Dark Helmet: Who is he?

Colonel Sandurz: He's an asshole sir.

Dark Helmet: I know that! What's his name?

Colonel Sandurz: That is his name sir. Asshole. Major Asshole!

Dark Helmet: And his cousin?

Colonel Sandurz: He's an asshole too sir. Gunner's mate First Class Philip Asshole!

Dark Helmet: How many assholes we got on this ship, anyhow?

[Entire bridge crew stands up]

Bridge Crew: Yo!

[Dark Helmet turns around, startled]

Dark Helmet: I knew it. I'm surrounded by assholes!

[Dark Helmet pulls down his face shield]

Dark Helmet: Keep firing, assholes!

- Mel Brooks (*Spaceballs*, 1987)

Chapter 8 Appendix

8.1 PCR primers

Table 8.1 PCR primers. The gene from which in which the target fragment sits is given in the first column.

Frag-ment	Primer name	HXB2 Location	Inner/ Outer	Sequence
gag	1F	627-645	Both	CTCTAGCAGTGGCGCCCGA
	2F	641-661	Inner	CCCGACACGGGACTTGAAAGC
	F1R	2045-2024	Inner	TCCTTTCCACATTTCCAACAGC
	GAGN1-5-OP	693-711	Outer	AGGACTCGGCTTGCTGAAG
	GAGN1-3-OP	1942-1921	Outer	TCCTAAAATTGCCTCTCTGCAT
	GAGN1-5-IP	768-787	Outer	AGCGGAGGCTAGAAGGAGAC
	GAGN1-3-IP	1903-1882	Outer	TTGTTACTTGGCTCATTGCTTC
	GagC5-OP	627-646	Outer	CTCTAGCAGTGGCGCCCGAA
	GagC3-OP	696-714	Outer	TCCTTTCCACATTTCCAACAGCC
	GagC5-IP	2045-2023	Inner	ACTCGGCTTGCTGAAGTGC
	GagC3-IP	2003-1984	Inner	CAATTTCTGGCTATGTGCCC
pol	HXB2-4F	1858-1880	Outer	AGCCATAAAGCAAGAGTTTTGGC
	HXB2-6R	4083-4060	Outer	GATCTGGTTGTGCTTGAATGTTTC
Frag-ment	Primer name	HXB2 Location	Inner/ Outer	Sequence
	HXB2-5F	2509-2533	Inner	GAAGAAATCTGTTGACTCAGATTGG
	HXB2-5R	3508-3489	Inner	TCATAATACACTCCATGTAC
env	VIF	4900-4923	Outer	GGGTTTATTACAGGGACAGCAGAG
	OFM19	9632-9603	Outer	GCACTCAAGGCAAGCTTTATTGAGGCTTA

	envA	5954-5982	Inner	GGCTTAGGGATCTCCTATGGCAGGAAGAA
	envN	9171-9144	Inner	CTGCAAATCAGGGAACTAGCCTTGTGT
	env5B-out	5853-5877	Outer	TAGAGCCCTGGAAGCATCCAGGAAG
	env3B-out	8936-8913	Outer	GCTACTTGTGATTGCTCCATGT
	env5B-in	5953-5983	Outer	CACCTTAGGCATCTCCTATGGCAGGAAGAAG
	env3B-in	8904-8882	Outer	GTCTCGAGATACTCGTCCCACCC
nef	NefN1-5-OP	8016-8039	Outer	TGCTCTGGAAAACTCATTTGCACC
	NefN1-3-OP	9570-9548	Outer	CAGGCTCAGATCTGGTCTAACCA
	NefN1-5-IP	8685-8703	Inner	GCACTAGCTGCGGGGACAG
	NefN1-3-IP	9528-9509	Inner	TACAGGCAAAAAGCAGCTGCTTATATG

8.2 Additional primers for Sanger sequencing

Table 8.2 Additional primers for Sanger sequencing

Frag-ment	Primer Name	HXB2 Location	Sequence
gag	GP-2F	1307-1328	CAGCATTATCAGAAGGAGCCAC
	gp1NS1	1737-1756	GATGACAGAAACCTTGTTGG
	GP-3F	1821-1840	AGAAATGATGACAGCATGTC
	GP-4R	1337-1316	TCTTGTGGGGGTGGCTCCTTCTG
	GP-3R	1507-1486	TTCCTGCTATGTCACTTCC
env	VES3	6550-6570	TCAGTTTATGGGATCAAAGCC
	E90	6953-6976	CACAGTACAATGTACACATGGAAT
	VES5	7060-7084	ACAATGCTAAAACCATAATAGTACA
	envS3	7780-7798	TCCTTGGGTTCTTGGGAGC
	E145R	7981-7003	AGCAGTTGAGTTGATACTACTGG
	VES4	7096-7119	GTCTTGTACAATTAATTTCTACAG
	envS4	8222-8241	ACCAATTCCACAAACTTGCC
nef	Nef5-OP	8999-9019	TTCCAGTCAGACCTCAGGTAC
	Nef5-IP	9085-9063	CTTTTTAAAAGAAAAGGGGGGAC

8.3 Poisson Fitter Histograms for Chapter 3

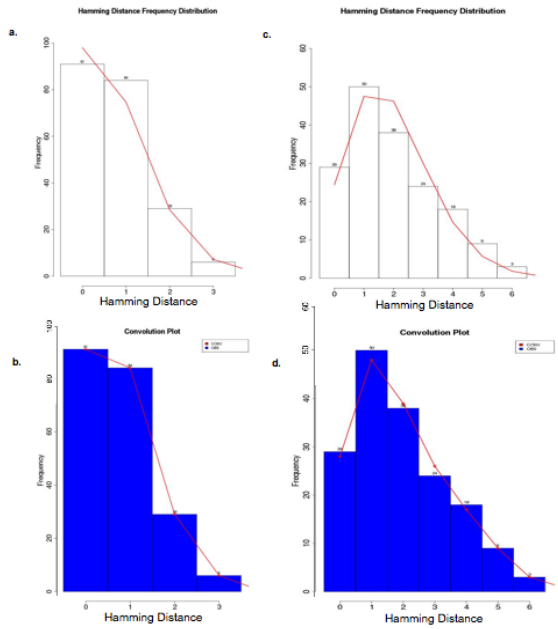


Figure 8.1 Histograms for the HD frequency distribution for the long* env *fragment in P1 (a. and b.) and P2 (c. and d.), with APOBEC signatures removed. The x-axis plots Hamming distance values from the within-host consensus for each recipient. The y-axis plots HD frequency counts as a percentage of samples sequenced. The red line in a. and c. plots the expected HD frequency counts under a Poisson distribution model. The red lines in b. and d. plot the expected HD frequency counts if the data followed a star-like phylogeny under the self-convolution model implemented in Poisson-Fitter.

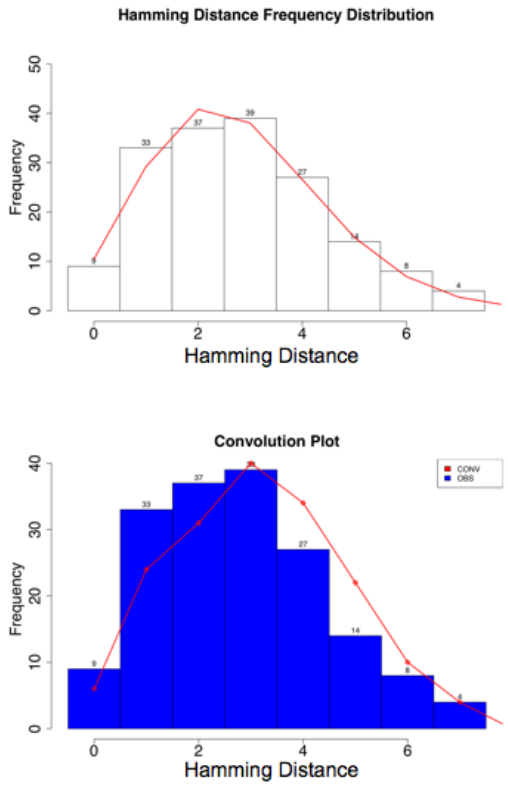


Figure 8.2 Histograms for the HD frequency distribution for the long* env *fragment in P2, with APOBEC signatures included. The x-axis plots Hamming distance values from the within-host consensus for each recipient. The y-axis plots HD frequency counts as a percentage of samples sequenced. The red line in the top plot is the expected HD frequency counts under a Poisson distribution model. The red lines in the bottom plot is the expected HD frequency counts if the data followed a star-like phylogeny under the self-convolution model implemented in Poisson-Fitter.

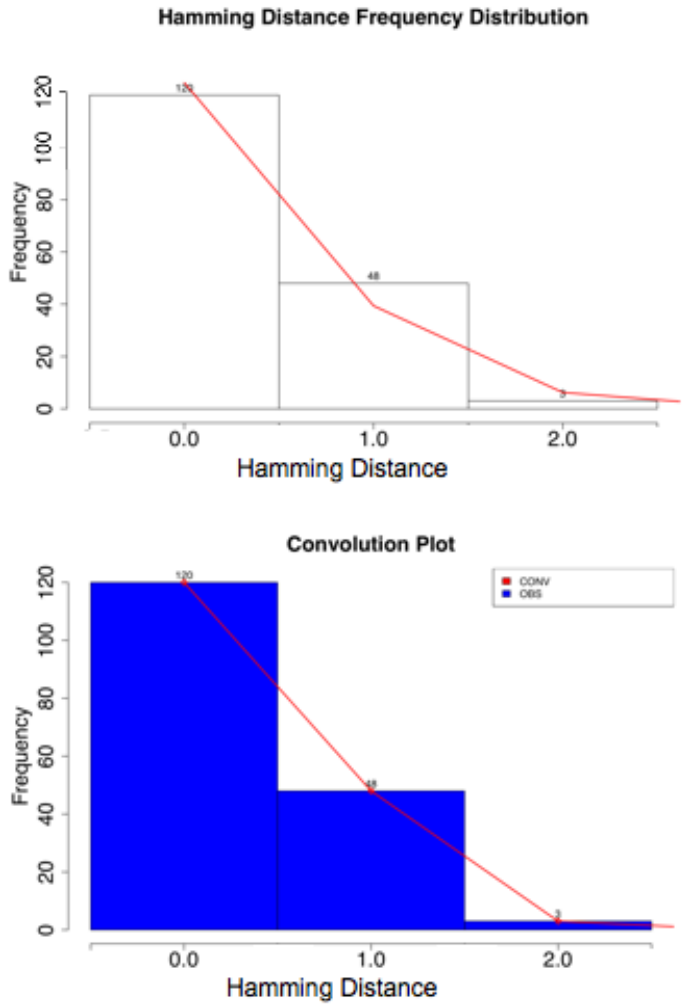


Figure 8.3 Histograms for the HD frequency distribution for the* env C2V5 *fragment in P2, with APOBEC signatures included. The x-axis plots Hamming distance values from the within-host consensus for each recipient. The y-axis plots HD frequency counts as a percentage of samples sequenced. The red line in the top plot is the expected HD frequency counts under a Poisson distribution model. The red lines in the bottom plot is the expected HD frequency counts if the data followed a star-like phylogeny under the self-convolution model implemented in Poisson-Fitter.

8.4 Neutralization and infectivity data for Chapter 3

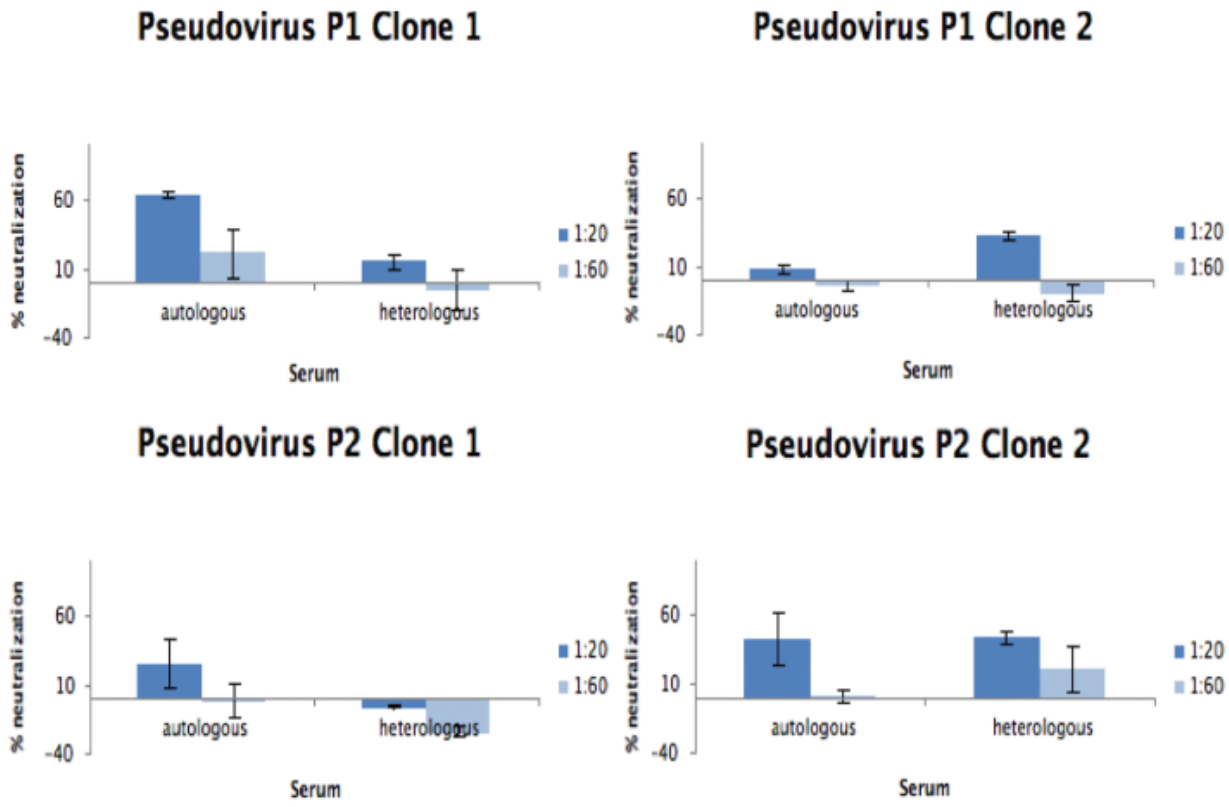


Figure 8.4 In vitro *neutralization of participant env-pseudotyped viruses by autologous (same participant) and heterologous (participant's partner) serum for each of the clones tested. Data from the most concentrated serum dilutions are shown, with the colour scheme provided by a legend at the right of each graph.*

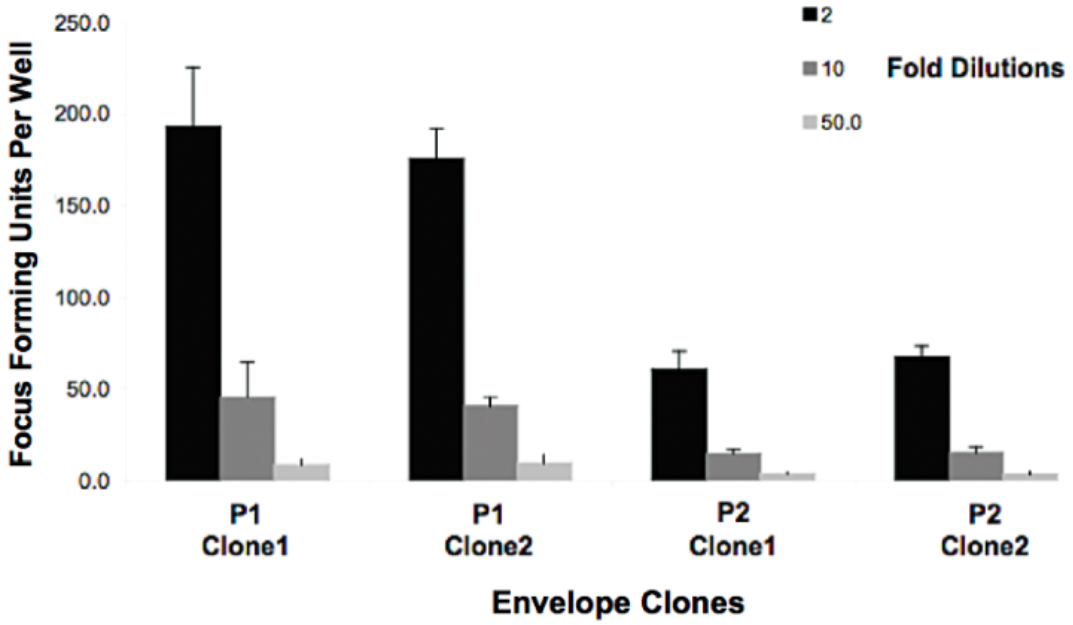


Figure 8.5 Infectivity of the env-pseudotyped virus from each participant. Data for different fold dilutions of pseudovirus after normalization to reverse transcriptase levels. A colour key indicates the dilutions tested.

8.5 Additional robust analysis for Chapter 4

8.5.1 Robust analysis in *env C2V5* using the TN93 model

8.5.1.1 Effect of using TN93 model without a mean rate prior

I re-analysed the data in the *C2V5 env* region using a less parameter-rich substitution model with the ULRM clock but without enforcing a strong substitution rate prior (Figure 8.6). I chose the TN93 model (Lemey et al. 2009; Tamura and Nei 1993) and convergence was achieved using the relaxed clock analysis with no mean rate prior.

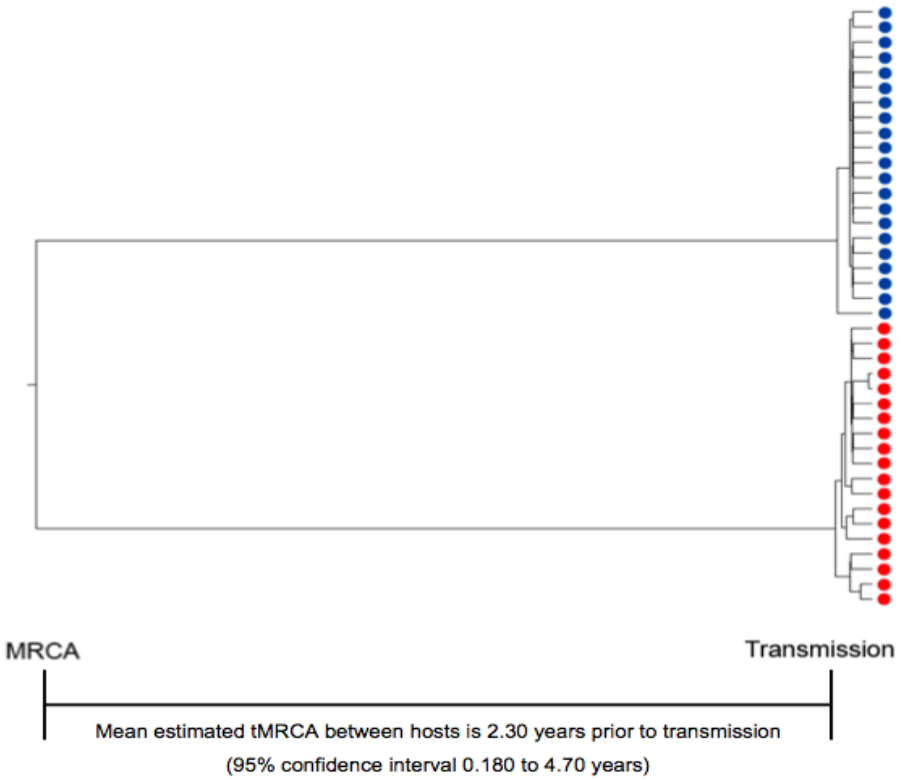


Figure 8.6* env C2V5*: TN93 model, ULRM clock, no mean rate prior; blue closed dots=P1, red closed dots=P2.

The estimate for the lower 95% CL was a lot less than one year prior to transmission (Figure 8.6, Table 4.2).

8.5.1.2 Using the TN93 model with the mean rate prior

I re-analysed the data with the normal prior of the mean rate

estimated from the independent dataset (Figure 8.7).

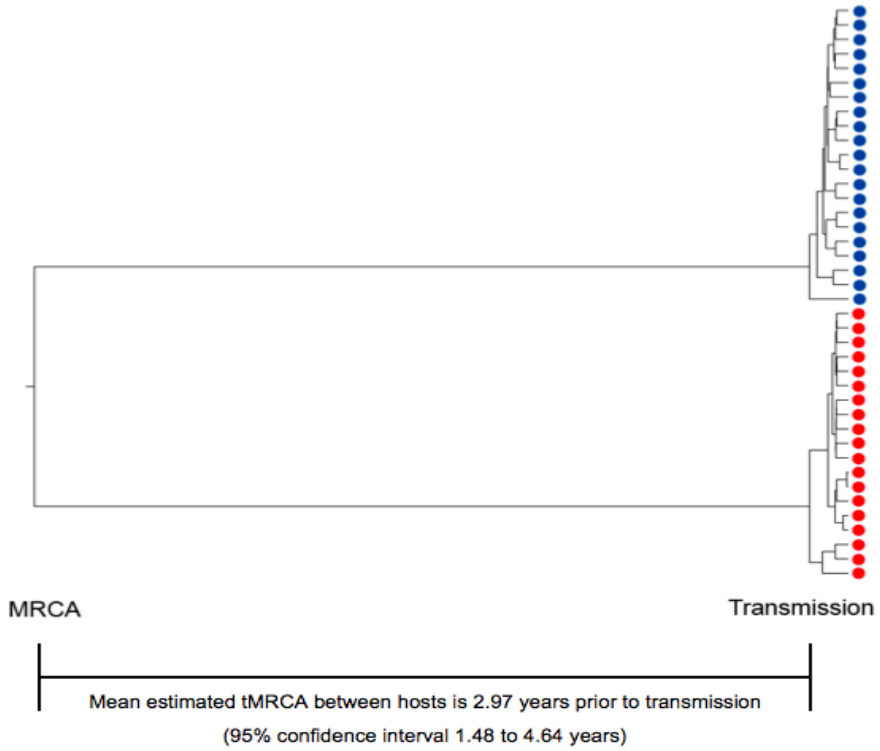


Figure 8.7* env C2V5*: TN93 model, uncorrelated lognormal molecular clock, normal mean rate prior; blue closed dots=P1, red closed dots=P2.

Thus I applied the same mean rate prior for a ULRM clock (Lemey et al. 2006). The mean rate prior restored the lower 95% CL for the estimated tMRCA to longer than one year prior to transmission (Figure 8.7, Table 4.2).

8.5.1.3 The effect of molecular clock choice

I repeated the analysis using the TN93 substitution model with a strict molecular clock and no substitution rate prior. The lower 95% CL was almost one year prior to transmission (Figure 8.8, Table 4.2).

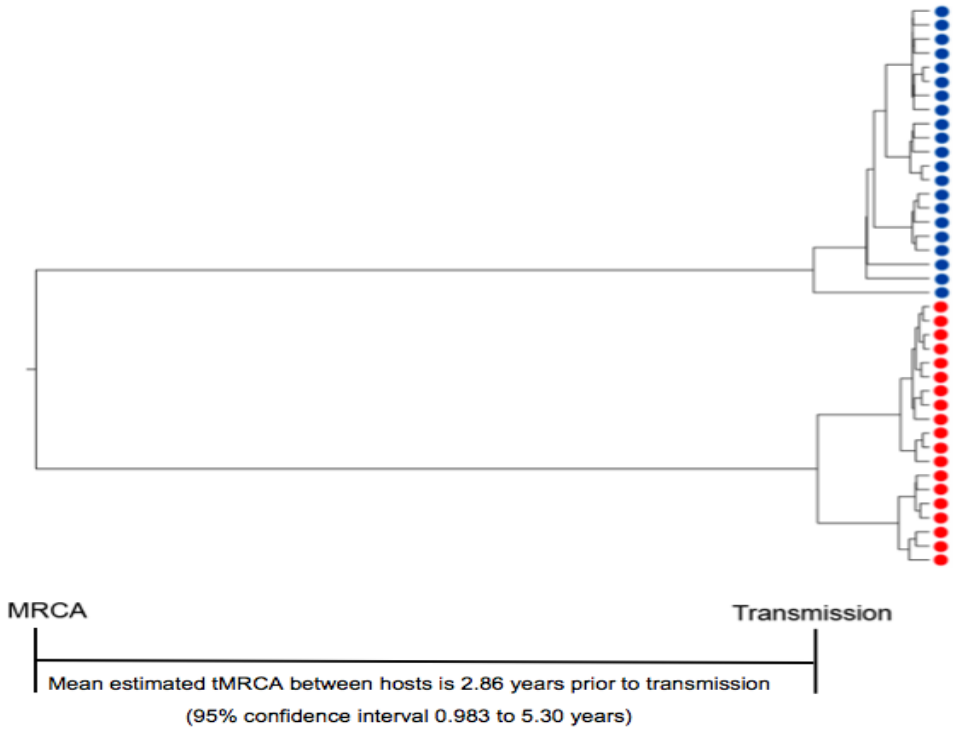
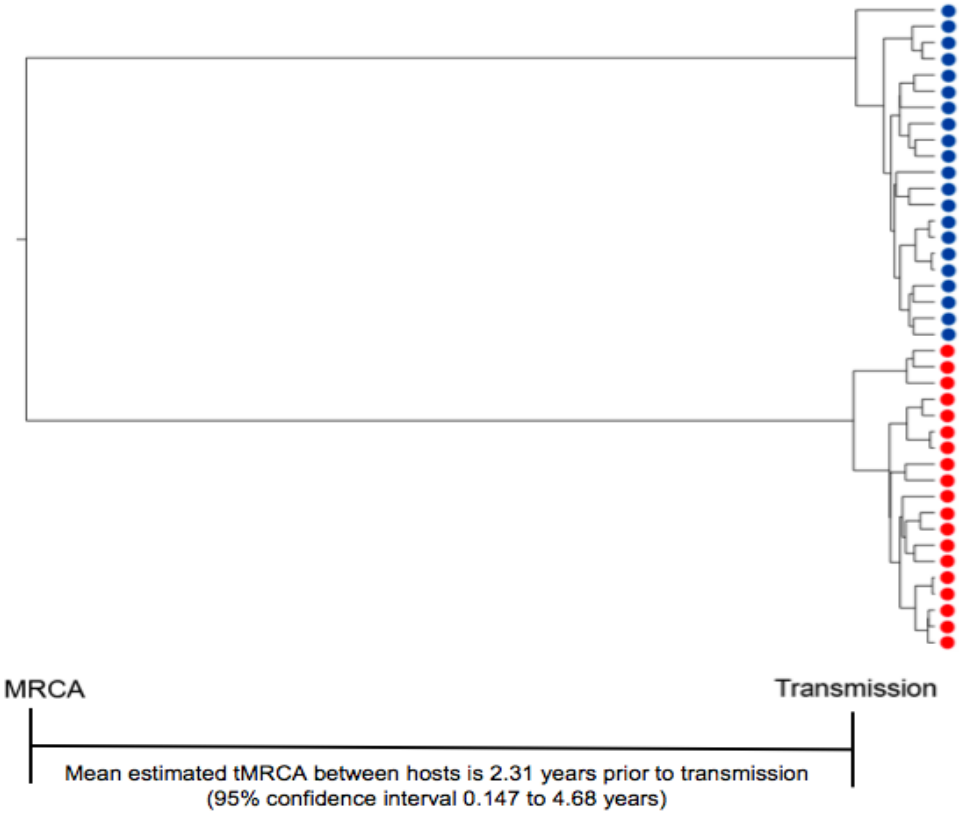


Figure 8.8* env C2V5*: TN93 model, strict molecular clock, no substitution rate prior; blue closed dots=P1, red closed dots=P2.

8.5.2 Robust analysis in *env C2V5* using the HKY85 model

To ensure that the lower 95% CL estimate for *env C2V5* was not significantly sensitive to the substitution model, I repeated my relaxed-clock analyses (Figures 8.9 and 8.10) using the nested HKY85 model. There was very little difference in the confidence interval estimates (Table 4.2).



8.9* env C2V5*: HKY85 model, ULRM clock, no mean rate prior; blue closed dots=P1, red closed dots=P2.

Once again, including the mean rate prior (Figure 8.10) increased the estimated interval between the lower 95% confidence interval for the between-host tMRCA and the time of transmission to greater than one year (Table 4.2).

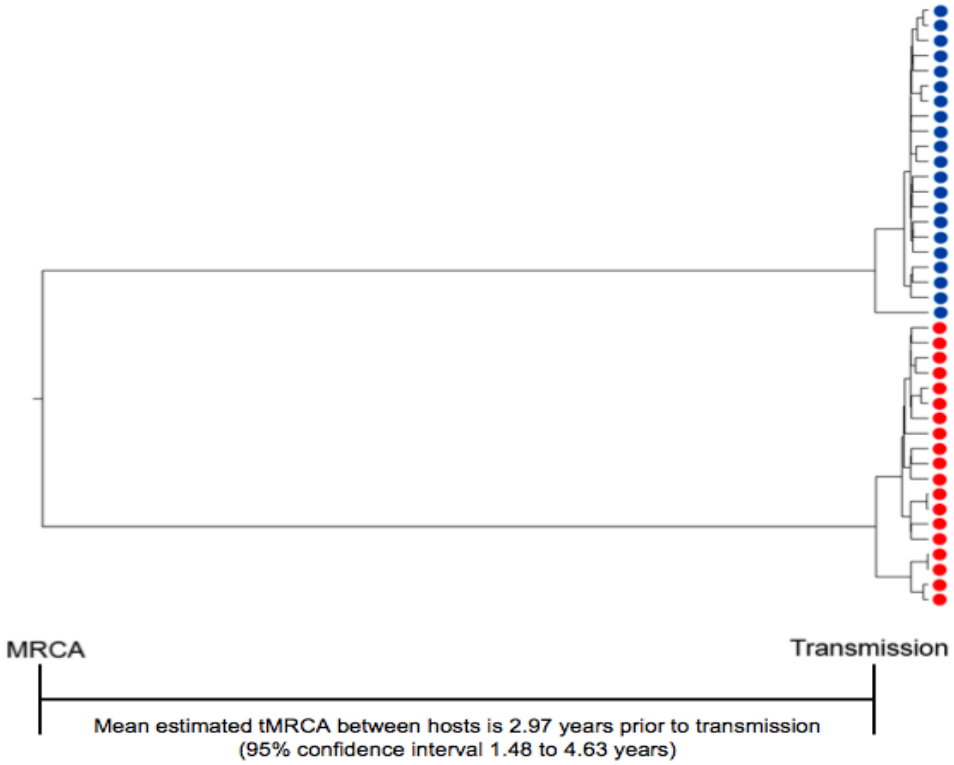


Figure 8.10* env C2V5*: HKY85 model, ULRM clock, normal mean rate prior; blue closed dots=P1, red closed dots=P2.

8.5.3 Rate heterogeneity among sites

Previous authors have assumed that *env gp120* evolves with virtual neutrality during primary infection (Keele et al. 2008; Lemey et al. 2006). For this analysis, α parameter for rate heterogeneity among sites was estimated from the data but made little difference to the between-host tMRCA estimate (Figure 8.11, Table 4.2).

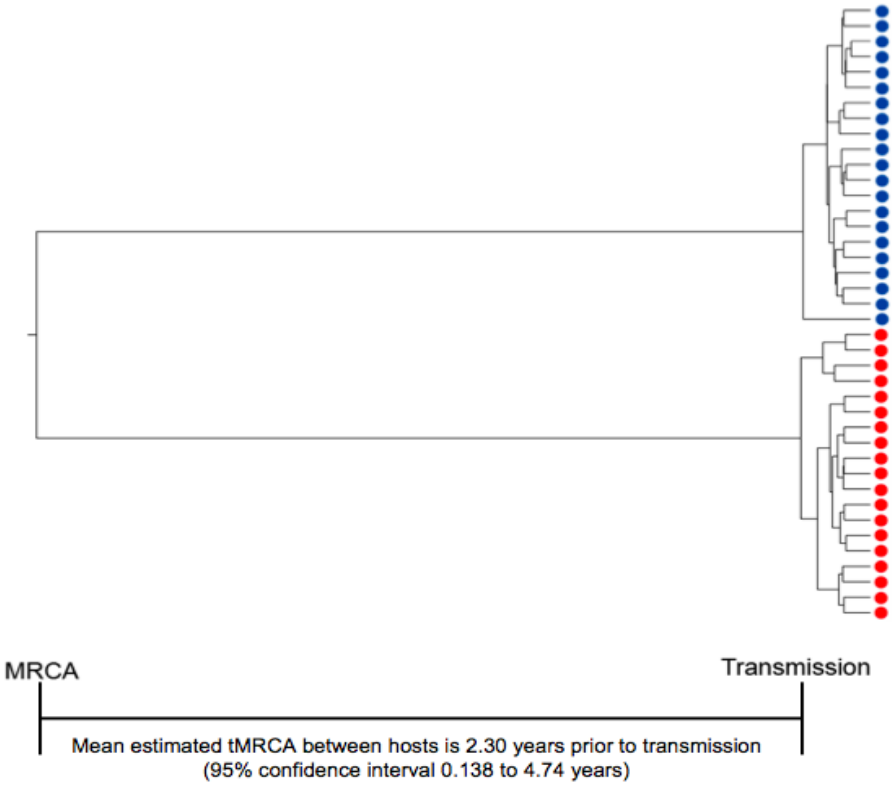


Figure 8.11 TN93+Γ model, ULRM clock, no mean rate prior; blue closed dots=P1, red closed dots=P2.

8.6 ELISpot data for SL9 mutation in Chapters 5 and 6

Table 8.3 Epitope sequences and interferon-γ ELISpot responses to SL9 wild-type peptide, SL9 variant peptides associated with escape from HLA-A*0201, and overlapping peptides (OLP) containing part or all of SL9. A positive response (+) was defined as greater than 50 spots/million cells (minus background spots observed in the negative control). The wild-type (WT) peptide is highlighted in bold black font. The SL9 Y79F mutant found in P1 from day 235 is highlighted in blue font. Amino acid substitutions away from the WT are highlighted in red, bold and underlined font.

Peptide		Time points		
Name	AA Sequence	Day 63	Day 235	Day 474
SL9 WT	**SLYNTVATL**	**negative**	**negative**	**negative**
SL9 variant 1	SL**F**N**AI**A**V**L	negative	negative	negative
SL9 variant 2	SL**F**N**A**VA**V**L	negative	negative	negative
SL9 variant 3	SLYNTVA**V**L	negative	100	negative
SL9 variant 4	SLYNT**I**A**V**L	negative	negative	negative
SL9 variant 5	SLYNT**I**ATL	negative	60	negative
SL9 variant 6	SL**F**NT**I**A**V**L	negative	negative	negative
SL9 variant 7	SL**F**NTVATL	negative	negative	negative
SL9 variant 8	SL**F**NTVA**V**L	negative	negative	negative
SL9 variant 9	SLYNTVAT**I**	negative	negative	negative
SL9 variant 10	SL**F**NTVAT**V**	negative	negative	negative
SL9 variant 11	SLYNT**ISV**L	negative	negative	negative
SL9 variant 12	SLYNT**IVV**L	negative	negative	negative
SL9 variant 13	SL**F**NT**I**ATL	negative	negative	negative
Gag OLP 17	QPSLQTGSEELR**SLY**	negative	negative	negative
Gag OLP 18	QTGSEELR**SLYNTVA**	negative	negative	negative
Gag OLP 19	EELR**SLYNTVATL**YC	negative	negative	negative
Gag OLP 20	**SLYNTVATL**YCVHQR	negative	negative	negative
Gag OLP 21	**TVATL**YCVHQRIEVK	negative	negative	negative
Gag OLP 22	**L**YCVHQRIEVKDTKE	negative	negative	negative

Chapter 9 Published paper

English, S. Katzourakis, A. Bonsall, D. Flanagan, P. Duda, A. Fidler, S. Weber, J. McClure, M. SPARTAC Trial Investigators, Phillips, R. Frater, J. (2011), 'Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants', *Retrovirology,* 8, 54.

ABSTRACT

Background To combat the pandemic of human immunodeficiency virus 1 (HIV-1), a successful vaccine will need to cope with the variability of transmissible viruses. Human hosts infected with HIV-1 potentially harbour many viral variants but very little is known about viruses that are likely to be transmitted, or even if there are viral characteristics that predict enhanced transmission *in vivo*. We show for the first time that genetic divergence consistent with a single transmission event *in vivo* can represent several years of pre-transmission evolution.

Results We describe a highly unusual case consistent with a single donor transmitting highly related but distinct HIV-1 variants to two individuals on the same evening. We confirm that the clustering of viral genetic sequences present within each recipient is consistent with the history of a single donor across the viral *env, gag* and *pol* genes by maximum likelihood and Bayesian Markov Chain Monte Carlo based phylogenetic analyses. Based on an uncorrelated, lognormal relaxed clock of *env* gene evolution calibrated with other datasets, the time since the most recent common ancestor is estimated as 2.86 years prior to transmission (95% confidence interval 1.28 to 4.54 years).

Conclusion Our results show that an effective design for a preventative vaccine will need to anticipate extensive HIV-1 diversity within an individual donor as well as diversity at the population level.

Chapter 10 References

Abbot, P., et al. (2011), 'Inclusive fitness theory and eusociality', *Nature,* 471 (7339), E1-4; author reply E9-10.

Abdool Karim, S. S., et al. (2011), 'Safety and effectiveness of BufferGel and 0.5% PRO2000 gel for the prevention of HIV infection in women', *AIDS,* 25 (7), 957-66.

Abrahams, M. R., et al. (2009), 'Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants', *J Virol,* 83 (8), 3556-67.

Agy, M. B., et al. (1992), 'Infection of Macaca nemestrina by human immunodeficiency virus type-1', *Science,* 257 (5066), 103-6.

Aiken, C., et al. (1994), 'Nef induces CD4 endocytosis: requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain', *Cell,* 76 (5), 853-64.

Albert, J., et al. (1990), 'Rapid development of isolate-specific neutralizing antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralization by autologous sera', *Aids,* 4 (2), 107-12.

Alter, G., et al. (2007a), 'Evolution of innate and adaptive effector cell functions during acute HIV-1 infection', *J Infect Dis,* 195 (10), 1452-60.

--- (2007b), 'Differential natural killer cell-mediated inhibition of HIV-1 replication based on distinct KIR/HLA subtypes', *J Exp Med,* 204 (12), 3027-36.

Altfeld, M., et al. (2002), 'HIV-1 superinfection despite broad CD8+ T-cell responses containing replication of the primary virus', *Nature,*

420 (6914), 434-9.

--- (2003), 'Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection', *Aids,* 17 (18), 2581-91.

Anderson, J. P., et al. (2000), 'Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E', *J Virol,* 74 (22), 10752-65.

Anderson, S., McClain, N., and Riviello, R. J. (2006), 'Genital findings of women after consensual and nonconsensual intercourse', *J Forensic Nurs,* 2 (2), 59-65.

Anonymous 'GapStreeze', <http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html>.

Anonymous 'PAUP*v4.0', <http://paup.csit.fsu.edu/>.

Anonymous 'Se-Al v2.0a11', <http://evolve.zoo.ox.ac.uk/software.html>.

Anonymous 'Los Alamos National Laboratory (LANL) HIV sequence database', <www.hiv.lanl.gov>.

Anonymous 'Recombination Identification Program (RIP) v3.0', <www.hiv.lanl.gov/content/sequence/RIP/RIP.html>.

Attia, S., et al. (2009), 'Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis', *AIDS,* 23 (11), 1397-404.

Baggaley, R. F., White, R. G., and Boily, M. C. (2010a), 'Infectiousness of HIV-infected homosexual men in the era of highly active antiretroviral therapy', *AIDS,* 24 (15), 2418-20.

--- (2010b), 'HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention', *Int J Epidemiol,* 39 (4), 1048-63.

Bailes, E., et al. (2003), 'Hybrid origin of SIV in chimpanzees', *Science,* 300 (5626), 1713.

Bar, K. J., et al. (2010), 'Wide variation in the multiplicity of HIV-1 infection among injection drug users', *J Virol,* 84 (12), 6241-7.

Barouch, D. H. (2008), 'Challenges in the development of an HIV-1 vaccine', *Nature,* 455 (7213), 613-9.

Barton, N. H. and Charlesworth, B. (1998), 'Why sex and recombination?', *Science,* 281 (5385), 1986-90.

Biggs, M., Stermac, L. E., and Divinsky, M. (1998), 'Genital injuries following sexual assault of women with and without prior sexual intercourse experience', *CMAJ,* 159 (1), 33-7.

Blankson, J. N., et al. (2007), 'Isolation and characterization of replication-competent human immunodeficiency virus type 1 from a subset of elite suppressors', *J Virol,* 81 (5), 2508-18.

Boeras, D. I., et al. (2011), 'Role of donor genital tract HIV-1 diversity in the transmission bottleneck', *Proc Natl Acad Sci U S A,* 108 (46), E1156-63.

Boerlijst, M., Bonhoeffer, S., and Nowak, M. A. (1996), 'Viral Quasi-Species and Recombination', *Proceedings of the Royal Society B,* 263, 1577-84.

Bonaparte, M. I. and Barker, E. (2004), 'Killing of human immunodeficiency virus-infected primary T-cell blasts by autologous natural killer cells is dependent on the ability of the virus to alter the expression of major histocompatibility complex class I molecules', *Blood,* 104 (7), 2087-94.

Brenchley, J. M., et al. (2004), 'CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract', *J Exp Med,* 200 (6), 749-59.

Brennan, G., Kozyrev, Y., and Hu, S. L. (2008), 'TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis', *Proc Natl Acad Sci U S A,* 105 (9), 3569-74.

Brennan, G., et al. (2007), 'Novel TRIM5 isoforms expressed by Macaca nemestrina', *J Virol,* 81 (22), 12210-7.

Brockman, M. A., et al. (2010), 'Early selection in Gag by protective HLA alleles contributes to reduced HIV-1 replication capacity that may be largely compensated for in chronic infection', *J Virol,* 84 (22), 11937-49.

Brown, R. J., et al. (2011), 'Intercompartmental recombination of HIV-1 contributes to env intrahost diversity and modulates viral tropism and sensitivity to entry inhibitors', *J Virol,* 85 (12), 6024-37.

Brumme, Z. L., et al. (2008a), 'Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection', *AIDS,* 22 (11), 1277-86.

--- (2008b), 'Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection', *J Virol,* 82 (18), 9216-27.

Bryant, D. and Moulton, V. (2004), 'Neighbor-net: an agglomerative method for the construction of phylogenetic networks', *Mol Biol Evol,* 21 (2), 255-65.

Buchbinder, S. P., et al. (2008), 'Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial', *Lancet,* 372 (9653), 1881-93.

Büchen-Osmond, C. (Ed) (2004), '00.061.1.06.001. Human immunodeficiency virus 1', *ICTVdB - The Universal Virus Database,*

version 3.

Bull, M., et al. (2009), 'Compartmentalization of HIV-1 within the female genital tract is due to monotypic and low-diversity variants not distinct viral populations', *PLoS One,* 4 (9), e7122.

Butler, D. M., et al. 'The origins of sexually transmitted HIV among men who have sex with men', *Sci Transl Med,* 2 (18), 18re1.

Carrington, M., et al. (1999), 'HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage', *Science,* 283 (5408), 1748-52.

Casado, C., et al. (2007), 'Coinfection and superinfection in patients with long-term, nonprogressive HIV-1 disease', *J Infect Dis,* 196 (6), 895-9.

Castro, E., et al. (2008), 'Independent evolution of hypervariable regions of HIV-1 gp120: V4 as a swarm of N-Linked glycosylation variants', *AIDS Res Hum Retroviruses,* 24 (1), 106-13.

Cella, M., et al. (2009), 'A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity', *Nature,* 457 (7230), 722-5.

Centers for Disease Control and Prevention, CDC (2012), 'HIV Transmission', (updated 25 March 2012) <http://www.cdc.gov/hiv/resources/qa/transmission.htm>, accessed 1 September 2012.

Centlivre, M., et al. (2005), 'HIV-1 clade promoters strongly influence spatial and temporal dynamics of viral replication in vivo', *J Clin Invest,* 115 (2), 348-58.

Chakrabarti, L., et al. (1987), 'Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses', *Nature,* 328 (6130), 543-7.

Chakrabarti, L. A. (2004), 'The paradox of simian immunodeficiency

virus infection in sooty mangabeys: active viral replication without disease progression', *Front Biosci,* 9, 521-39.

Choe, H., et al. (1996), 'The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates', *Cell,* 85 (7), 1135-48.

Chohan, B., et al. (2005), 'Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1-V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels', *J Virol,* 79 (10), 6528-31.

Clamp, M., et al. (2004), 'The Jalview Java alignment editor', *Bioinformatics,* 20 (3), 426-7.

Clark, S. J. and Shaw, G. M. (1993), 'The acute retroviral syndrome and the pathogenesis of HIV-1 infection', *Semin Immunol,* 5 (3), 149-55.

Clavel, F., et al. (1989), 'Genetic recombination of human immunodeficiency virus', *J Virol,* 63 (3), 1455-9.

Clerici, M., et al. (1992), 'Cell-mediated immune response to human immunodeficiency virus (HIV) type 1 in seronegative homosexual men with recent sexual exposure to HIV-1', *J Infect Dis,* 165 (6), 1012-9.

Clever, J., Sassetti, C., and Parslow, T. G. (1995), 'RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1', *J Virol,* 69 (4), 2101-9.

Cohen, G. B., et al. (1999), 'The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells', *Immunity,* 10 (6), 661-71.

Cohen, O. J., Kinter, A., and Fauci, A. S. (1997), 'Host factors in the

pathogenesis of HIV disease', *Immunol Rev,* 159, 31-48.

Colonna, M., et al. (1993), 'HLA-C is the inhibitory ligand that determines dominant resistance to lysis by NK1- and NK2-specific natural killer cells', *Proc Natl Acad Sci U S A,* 90 (24), 12000-4.

Connor, R. I., et al. (1997), 'Change in coreceptor use correlates with disease progression in HIV-1--infected individuals', *J Exp Med,* 185 (4), 621-8.

Coplan, P. M., et al. (1996), 'Human immunodeficiency virus infection in Mexico City. Rectal bleeding and anal warts as risk factors among men reporting sex with men', *Am J Epidemiol,* 144 (9), 817-27.

Craig, H. M., Pandori, M. W., and Guatelli, J. C. (1998), 'Interaction of HIV-1 Nef with the cellular dileucine-based sorting pathway is required for CD4 down-regulation and optimal viral infectivity', *Proc Natl Acad Sci U S A,* 95 (19), 11229-34.

Crow, J. F. and Kimura, M. (1965), 'Evolution in sexual and asexual populations ', *Am Nat,* 99, 439-50.

Dalmau, J., et al. (2012), 'In-depth characterization of viral isolates from plasma and cells compared with plasma circulating quasispecies in early HIV-1 infection', *PLoS One,* 7 (2), e32714.

Darwin, Charles (1859), *On the origin of species by means of natural selection* (London,: J. Murray) ix, 1 , 502 p.

Deacon, N. J., et al. (1995), 'Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients', *Science,* 270 (5238), 988-91.

Dean, M., et al. (1996), 'Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco

City Cohort, ALIVE Study', *Science,* 273 (5283), 1856-62.

Delwart, E. L., et al. (1994), 'Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays', *J Virol,* 68 (10), 6672-83.

--- (1998), 'Human immunodeficiency virus type 1 populations in blood and semen', *J Virol,* 72 (1), 617-23.

Deng, H., et al. (1996), 'Identification of a major co-receptor for primary isolates of HIV-1', *Nature,* 381 (6584), 661-6.

Denton, P. W., et al. (2011), 'One percent tenofovir applied topically to humanized BLT mice and used according to the CAPRISA 004 experimental design demonstrates partial protection from vaginal HIV infection, validating the BLT model for evaluation of new microbicide candidates', *J Virol,* 85 (15), 7582-93.

Derdeyn, C. A. and Hunter, E. (2008), 'Viral characteristics of transmitted HIV', *Curr Opin HIV AIDS,* 3 (1), 16-21.

Derdeyn, C. A., et al. (2004), 'Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission', *Science,* 303 (5666), 2019-22.

Desrosiers, R. C. (1990), 'The simian immunodeficiency viruses', *Annu Rev Immunol,* 8, 557-78.

Diem, K., et al. (2008), 'Male genital tract compartmentalization of human immunodeficiency virus type 1 (HIV)', *AIDS Res Hum Retroviruses,* 24 (4), 561-71.

Drummond, A. J. and Rambaut, A. (2007), 'BEAST: Bayesian evolutionary analysis by sampling trees', *BMC Evol Biol,* 7, 214.

Drummond, A. J., et al. (2002), 'Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data', *Genetics,* 161 (3), 1307-20.

--- (2005), 'Bayesian coalescent inference of past population dynamics from molecular sequences', *Mol Biol Evol,* 22 (5), 1185-92.

--- (2006), 'Relaxed phylogenetics and dating with confidence', *PLoS Biol,* 4 (5), e88.

Duda, A., et al. (2009), 'HLA-associated clinical progression correlates with epitope reversion rates in early human immunodeficiency virus infection', *J Virol,* 83 (3), 1228-39.

Duerr, A., et al. (2012), 'Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study)', *J Infect Dis,* 206 (2), 258-66.

Edwards, C. T., et al. (2006), 'Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1', *BMC Evol Biol,* 6, 28.

Engel, G.A., et al. (2012), 'Naturally acquired Mycobacterium tuberculosis complex in laboratory pig-tailed macaques', *Emer Micro & Infect,* 1.

Epstein, L. G., et al. (1985), 'Progressive encephalopathy in children with acquired immune deficiency syndrome', *Ann Neurol,* 17 (5), 488-96.

Erlich, H. A., Gelfand, D., and Sninsky, J. J. (1991), 'Recent advances in the polymerase chain reaction', *Science,* 252 (5013), 1643-51.

Etienne, L., et al. (2011), 'Characterization of a new simian immunodeficiency virus strain in a naturally infected Pan troglodytes troglodytes chimpanzee with AIDS related symptoms', *Retrovirology,* 8, 4.

Fauci, A. S. and Fischinger, P. J. (1988), 'The development of an AIDS vaccine: progress and promise', *Public Health Rep,* 103 (3), 230-6.

Fauci, A. S., Mavilio, D., and Kottilil, S. (2005), 'NK cells in HIV infection: paradigm for protection or targets for ambush', *Nat Rev Immunol,* 5 (11), 835-43.

Feldman, M. W., Otto, S. P., and Christiansen, F. B. (1996), 'Population genetic perspectives on the evolution of recombination', *Annu Rev Genet,* 30, 261-95.

Fenizia, C., et al. (2011), 'TRIM5alpha does not affect simian immunodeficiency virus SIV(mac251) replication in vaccinated or unvaccinated Indian rhesus macaques following intrarectal challenge exposure', *J Virol,* 85 (23), 12399-409.

Fidler, S., et al. (2002), 'Virological and immunological effects of short-course antiretroviral therapy in primary HIV infection', *Aids,* 16 (15), 2049-54.

Fiebig, E. W., et al. (2005), 'Intermittent low-level viremia in very early primary HIV-1 infection', *J Acquir Immune Defic Syndr,* 39 (2), 133-7.

--- (2003), 'Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection', *Aids,* 17 (13), 1871-9.

Finzi, A., et al. (2012), 'Lineage-specific differences between human and simian immunodeficiency virus regulation of gp120 trimer association and CD4 binding', *J Virol,* 86 (17), 8974-86.

Fischer, W., et al. (2010), 'Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing', *PLoS One,* 5 (8), e12303.

Fisher, R. A. (1922), 'Darwinian evolution of mutations', *Eugen Rev,* 14 (1), 31-4.

--- (1930), *The Genetical Theory of Natural Selection* (London: Oxford

University Press).

Frahm, N., et al. (2004), 'Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities', *J Virol,* 78 (5), 2187-200.

Frank, S. A. (1994), 'Kin selection and virulence in the evolution of protocells and parasites', *Proc Biol Sci,* 258 (1352), 153-61.

--- (1996), 'Models of parasite virulence', *Q Rev Biol,* 71 (1), 37-78.

Frater, A. J., et al. (2007), 'Effective T-cell responses select human immunodeficiency virus mutants and slow disease progression', *J Virol,* 81 (12), 6742-51.

Froissart, R., et al. (2004), 'Co-infection weakens selection against epistatic mutations in RNA viruses', *Genetics,* 168 (1), 9-19.

Frost, S. D., et al. (2005), 'Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B', *J Virol,* 79 (10), 6523-7.

Fuentes, G. M., et al. (1996), 'Strand displacement synthesis of the long terminal repeats by HIV reverse transcriptase', *J Biol Chem,* 271 (4), 1966-71.

Gajdusek, D. C., et al. (1985), 'Infection of chimpanzees by human T-lymphotropic retroviruses in brain and other tissues from AIDS patients', *Lancet,* 1 (8419), 55-6.

Gantt, S., et al. (2010), 'Genetic analyses of HIV-1 env sequences demonstrate limited compartmentalization in breast milk and suggest viral replication within the breast that increases with mastitis', *J Virol,* 84 (20), 10812-9.

Gao, F., et al. (1999), 'Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes', *Nature,* 397 (6718), 436-41.

Gardner, A. and Grafen, A. (2009), 'Capturing the superorganism: a formal theory of group adaptation', *J Evol Biol,* 22 (4), 659-71.

Gaschen, B., et al. (2002), 'Diversity considerations in HIV-1 vaccine selection', *Science,* 296 (5577), 2354-60.

Gazzard, B. G., et al. (2008), 'British HIV Association Guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008', *HIV Med,* 9 (8), 563-608.

Georges-Courbot, M. C., et al. (1998), 'Natural infection of a household pet red-capped mangabey (Cercocebus torquatus torquatus) with a new simian immunodeficiency virus', *J Virol,* 72 (1), 600-8.

Giorgi, E. E., et al. (2010), 'Estimating time since infection in early homogeneous HIV-1 samples using a poisson model', *BMC Bioinformatics,* 11, 532.

Giorgi, J. V., et al. (1999), 'Shorter survival in advanced human immunodeficiency virus type 1 infection is more closely associated with T lymphocyte activation than with plasma virus burden or virus chemokine coreceptor usage', *J Infect Dis,* 179 (4), 859-70.

Goonetilleke, N., et al. (2009), 'The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection', *J Exp Med,* 206 (6), 1253-72.

Gordon, S. N., et al. (2007), 'Severe depletion of mucosal CD4+ T cells in AIDS-free simian immunodeficiency virus-infected sooty mangabeys', *J Immunol,* 179 (5), 3026-34.

Gottlieb, G. S., et al. (2004), 'Dual HIV-1 infection associated with rapid disease progression', *Lancet,* 363 (9409), 619-22.

Grabar, S., et al. (2009), 'Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the

French Hospital Database on HIV', *AIDS,* 23 (9), 1163-9.

Gray, E. S., et al. (2007a), 'Neutralizing antibody responses in acute human immunodeficiency virus type 1 subtype C infection', *J Virol,* 81 (12), 6187-96.

Gray, L., et al. (2007b), 'Phenotype and envelope gene diversity of nef-deleted HIV-1 isolated from long-term survivors infected from a single source', *Virol J,* 4, 75.

Gray, R. H., et al. (2007c), 'The impact of male circumcision on HIV incidence and cost per infection prevented: a stochastic simulation model from Rakai, Uganda', *Aids,* 21 (7), 845-50.

--- (2007d), 'Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial', *Lancet,* 369 (9562), 657-66.

Guadalupe, M., et al. (2003), 'Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy', *J Virol,* 77 (21), 11708-17.

Guindon, S. and Gascuel, O. (2003), 'A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood', *Syst Biol,* 52 (5), 696-704.

Guttman, D. S. and Dykhuizen, D. E. (1994), 'Clonal divergence in Escherichia coli as a result of recombination, not mutation', *Science,* 266 (5189), 1380-3.

Haaland, R. E., et al. (2009), 'Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1', *PLoS Pathog,* 5 (1), e1000274.

Haas, G., et al. (1996), 'Dynamics of viral variants in HIV-1 Nef and specific cytotoxic T lymphocytes in vivo', *J Immunol,* 157 (9), 4212-21.

Haase, A. T. (2005), 'Perils at mucosal front lines for HIV and SIV and their hosts', *Nat Rev Immunol,* 5 (10), 783-92.

Haldane, J. B. S. (1932), *The Causes of Evolution* (Princeton: Princeton University Press) 222.

Hamilton, W. D. (1964), 'The genetical evolution of social behaviour. I.', *J Theor Biol* 7, 1-16.

Hamlyn, E., et al. (2012), 'Increased levels of CD4 T-cell activation in individuals with CXCR4 using viruses in primary HIV-1 infection', *AIDS,* 26 (7), 887-90.

Hansen, S. G., et al. (2009), 'Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge', *Nat Med,* 15 (3), 293-9.

Harel-Bellan, A., et al. (1986), 'Natural killer susceptibility of human cells may be regulated by genes in the HLA region on chromosome 6', *Proc Natl Acad Sci U S A,* 83 (15), 5688-92.

Hasegawa, M., Kishino, H., and Yano, T. (1985), 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA', *J Mol Evol,* 22 (2), 160-74.

Hayman, A., et al. (2001), 'Phylogenetic analysis of multiple heterosexual transmission events involving subtype b of HIV type 1', *AIDS Res Hum Retroviruses,* 17 (8), 689-95.

Heath, L., et al. (2010), 'Restriction of HIV-1 genotypes in breast milk does not account for the population transmission genetic bottleneck that occurs following transmission', *PLoS One,* 5 (4), e10213.

Highleyman, L. (1999), 'Detuned assay used to track recent infections', *BETA,* 12 (4), 6, 78.

Hillis, D. M. and Huelsenbeck, J. P. (1992), 'Signal, noise, and reliability in molecular phylogenetic analyses', *J Hered,* 83 (3), 189-

95.

Hirsch, V. M. (2004), 'What can natural infection of African monkeys with simian immunodeficiency virus tell us about the pathogenesis of AIDS?', *AIDS Rev,* 6 (1), 40-53.

Hollingsworth, T. D., Anderson, R. M., and Fraser, C. (2008), 'HIV-1 transmission, by stage of infection', *J Infect Dis,* 198 (5), 687-93.

Holmes, E. C. (2009), *The Evolution and Emergence of RNA Viruses*, eds P.H. Harvey and R.M. May (Oxford Series in Ecology and Evolution: Oxford University Press).

Hu, J., Gardner, M. B., and Miller, C. J. (2000), 'Simian immunodeficiency virus rapidly penetrates the cervicovaginal mucosa after intravaginal inoculation and infects intraepithelial dendritic cells', *J Virol,* 74 (13), 6087-95.

Hu, W. S. and Temin, H. M. (1990a), 'Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination', *Proc Natl Acad Sci U S A,* 87 (4), 1556-60.

--- (1990b), 'Retroviral recombination and reverse transcription', *Science,* 250 (4985), 1227-33.

Huang, K. H., et al. (2011), 'Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations', *PLoS One,* 6 (4), e19018.

Huber, H. E., et al. (1989), 'Human immunodeficiency virus 1 reverse transcriptase. Template binding, processivity, strand displacement synthesis, and template switching', *J Biol Chem,* 264 (8), 4669-78.

Hue, S., et al. (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.

Huelsenbeck, J. P. and Ronquist, F. (2001), 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics,* 17 (8), 754-5.

Humes, D. and Overbaugh, J. (2011), 'Adaptation of subtype a human immunodeficiency virus type 1 envelope to pig-tailed macaque cells', *J Virol,* 85 (9), 4409-20.

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

Ignatius, R., et al. (2002), 'Increased macrophage infection upon subcutaneous inoculation of rhesus macaques with simian immunodeficiency virus-loaded dendritic cells or T cells but not with cell-free virus', *J Virol,* 76 (19), 9787-97.

Jordan, M. R., et al. (2010), 'Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations', *J Virol Methods,* 168 (1-2), 114-20.

Jung, A., et al. (2002), 'Recombination: Multiply infected spleen cells in HIV patients', *Nature,* 418 (6894), 144.

Karim, Q. A., et al. (2011), 'Recruitment of high risk women for HIV prevention trials: baseline HIV prevalence and sexual behavior in the CAPRISA 004 tenofovir gel trial', *Trials,* 12, 67.

Karlsson, A. C., et al. (2007), 'Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape', *PLoS One,* 2 (2), e225.

Kaslow, R. A., et al. (1990), 'A1, Cw7, B8, DR3 HLA antigen combination associated with rapid decline of T-helper lymphocytes in HIV-1 infection. A report from the Multicenter AIDS Cohort Study', *Lancet,* 335 (8695), 927-30.

Kaufmann, D. E., et al. (2004), 'Limited durability of viral control following treated acute HIV infection', *PLoS Med,* 1 (2), e36.

Kawashima, Y., et al. (2009), 'Adaptation of HIV-1 to human leukocyte antigen class I', *Nature,* 458 (7238), 641-5.

Kearney, M., et al. (2009), 'Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals', *J Virol,* 83 (6), 2715-27.

Keele, B. F., et al. (2009a), 'Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1', *J Exp Med,* 206 (5), 1117-34.

--- (2006), 'Chimpanzee reservoirs of pandemic and nonpandemic HIV-1', *Science,* 313 (5786), 523-6.

--- (2009b), 'Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz', *Nature,* 460 (7254), 515-9.

--- (2008), 'Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection', *Proc Natl Acad Sci U S A,* 105 (21), 7552-7.

Kemal, K. S., et al. (2003), 'HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation', *Proc Natl Acad Sci U S A,* 100 (22), 12972-7.

Kerkau, T., et al. (1992), 'Mechanism of MHC class I downregulation in HIV infected cells', *Immunobiology,* 184 (4-5), 402-9.

Kestler, H., et al. (1990), 'Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus', *Science,* 248 (4959), 1109-12.

Kiepiela, P., et al. (2007), 'CD8+ T-cell responses to different HIV proteins have discordant associations with viral load', *Nat Med,* 13 (1), 46-53.

Kimura, M. (1968), 'Genetic variability maintained in a finite population due to mutational production of neutral and nearly

neutral isoalleles', *Genet Res,* 11 (3), 247-69.

--- (1969), 'The rate of molecular evolution considered from the standpoint of population genetics', *Proc Natl Acad Sci U S A,* 63 (4), 1181-8.

--- (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *J Mol Evol,* 16 (2), 111-20.

Kimura, M. and Crow, J. F. (1964), 'The Number of Alleles That Can Be Maintained in a Finite Population', *Genetics,* 49, 725-38.

Kirchhoff, F., et al. (1995), 'Brief report: absence of intact nef sequences in a long-term survivor with nonprogressive HIV-1 infection', *N Engl J Med,* 332 (4), 228-32.

Kirmaier, A., et al. (2010), 'TRIM5 suppresses cross-species transmission of a primate immunodeficiency virus and selects for emergence of resistant variants in the new species', *PLoS Biol,* 8 (8).

Kitahata, M. M., et al. (2009), 'Effect of early versus deferred antiretroviral therapy for HIV on survival', *N Engl J Med,* 360 (18), 1815-26.

Kobayashi, T., et al. (2011), 'Identification of amino acids in the human tetherin transmembrane domain responsible for HIV-1 Vpu interaction and susceptibility', *J Virol,* 85 (2), 932-45.

Koelsch, K. K., et al. (2003), 'Clade B HIV-1 superinfection with wild-type virus after primary infection with drug-resistant clade B virus', *AIDS,* 17 (7), F11-6.

Korber, B., et al. (2000), 'Timing the ancestor of the HIV-1 pandemic strains', *Science,* 288 (5472), 1789-96.

Korber, B. T., et al. (1994), 'Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus

type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences', *J Virol,* 68 (11), 7467-81.

Kothe, D., et al. (2003), 'Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence', *J Acquir Immune Defic Syndr,* 33 (5), 625-34.

Kratovac, Z., et al. (2008), 'Primate lentivirus capsid sensitivity to TRIM5 proteins', *J Virol,* 82 (13), 6772-7.

Laboratory, Los Alamos National 'Highlighter', < www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html>.

Lamers, S. L., et al. (2009), 'Extensive HIV-1 intra-host recombination is common in tissues with abnormal histopathology', *PLoS One,* 4 (3), e5065.

Langlade-Demoyen, P., et al. (1994), 'Human immunodeficiency virus (HIV) nef-specific cytotoxic T lymphocytes in noninfected heterosexual contact of HIV-infected patients', *J Clin Invest,* 93 (3), 1293-7.

Lassen, K. G., et al. (2009), 'Elite suppressor-derived HIV-1 envelope glycoproteins exhibit reduced entry efficiency and kinetics', *PLoS Pathog,* 5 (4), e1000377.

Lecossier, D., et al. (2003), 'Hypermutation of HIV-1 DNA in the absence of the Vif protein', *Science,* 300 (5622), 1112.

Lee, B., et al. (1999), 'Quantification of CD4, CCR5, and CXCR4 levels on lymphocyte subsets, dendritic cells, and differentially conditioned monocyte-derived macrophages', *Proc Natl Acad Sci U S A,* 96 (9), 5215-20.

Lee, H. Y., et al. (2009), 'Modeling sequence evolution in acute HIV-1 infection', *J Theor Biol,* 261 (2), 341-60.

Leigh, E.G. (1990), 'Afterword', *The Causes of Evolution* (Princeton University Press).

Lemey, P. (2011).

Lemey, P., Rambaut, A., and Pybus, O. G. (2006), 'HIV evolutionary dynamics within and among hosts', *AIDS Rev,* 8 (3), 125-40.

Lemey, P., Salemi, M., and Vandamme, A. (2009), The Phylogenetic Handbook: A practical approach to phylogenetic analysis and hypothesis testing (2nd edn.: Cambridge University Press).

Lemey, P., et al. (2004), 'The molecular population genetics of HIV-1 group O', *Genetics,* 167 (3), 1059-68.

Letvin, N. L. (2006), 'Progress and obstacles in the development of an AIDS vaccine', *Nat Rev Immunol,* 6 (12), 930-9.

Letvin, N. L., et al. (1985), 'Induction of AIDS-like disease in macaque monkeys with T-cell tropic retrovirus STLV-III', *Science,* 230 (4721), 71-3.

--- (2007), 'No evidence for consistent virus-specific immunity in simian immunodeficiency virus-exposed, uninfected rhesus monkeys', *J Virol,* 81 (22), 12368-74.

Levy, D. N., et al. (2004), 'Dynamics of HIV-1 recombination in its natural target cells', *Proc Natl Acad Sci U S A,* 101 (12), 4204-9.

Levy, Jay A. (2007), *HIV and the pathogenesis of AIDS* (3rd edn.; Washington, D.C.: ASM Press) xvii, 644 p., 16 p. of plates.

Li, B., et al. (2006a), 'Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1', *J Virol,* 80 (11), 5211-8.

Li, H., et al. (2010), 'High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men', *PLoS Pathog,* 6 (5), e1000890.

Li, M., et al. (2005a), 'Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies', *J Virol,* 79 (16), 10108-25.

--- (2006b), 'Genetic and neutralization properties of subtype C human immunodeficiency virus type 1 molecular env clones from acute and early heterosexually acquired infections in Southern Africa', *J Virol,* 80 (23), 11776-90.

Li, Q., et al. (2005b), 'Peak SIV replication in resting memory CD4+ T cells depletes gut lamina propria CD4+ T cells', *Nature,* 434 (7037), 1148-52.

Lingappa, J. R., et al. (2010), 'Estimating the impact of plasma HIV-1 RNA reductions on heterosexual HIV-1 transmission risk', *PLoS One,* 5 (9), e12598.

Liu, R., et al. (1996), 'Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection', *Cell,* 86 (3), 367-77.

Liu, S. L., et al. (1997), 'Divergent patterns of progression to AIDS after infection from the same source: human immunodeficiency virus type 1 evolution and antiviral responses', *J Virol,* 71 (6), 4284-95.

Liu, Y., et al. (2008), 'Env length and N-linked glycosylation following transmission of human immunodeficiency virus Type 1 subtype B viruses', *Virology,* 374 (2), 229-33.

Lobritz, M. A., et al. (2007), 'Natural variation in the V3 crown of human immunodeficiency virus type 1 affects replicative fitness and entry inhibitor sensitivity', *J Virol,* 81 (15), 8258-69.

Lole, K. S., et al. (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), 152-60.

Low, A. J., et al. (2007), 'Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates', *AIDS,* 21 (14), F17-24.

Magiorkinis, G., et al. (2003), 'In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity', *J Gen Virol,* 84 (Pt 10), 2715-22.

Mansky, L. M. and Temin, H. M. (1995), 'Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase', *J Virol,* 69 (8), 5087-94.

Mao-Jones, J., et al. (2010), 'How microbial community composition regulates coral disease development', *PLoS Biol,* 8 (3), e1000345.

Mariani, R. and Skowronski, J. (1993), 'CD4 down-regulation by nef alleles isolated from human immunodeficiency virus type 1-infected individuals', *Proc Natl Acad Sci U S A,* 90 (12), 5549-53.

Markowitz, M., et al. (2003), 'A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo', *J Virol,* 77 (8), 5037-8.

Mascola, J. R., et al. (2005), 'Recommendations for the design and use of standard virus panels to assess neutralizing antibody responses elicited by candidate human immunodeficiency virus type 1 vaccines', *J Virol,* 79 (16), 10103-7.

--- (2000), 'Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies', *Nat Med,* 6 (2), 207-10.

--- (1996), 'Immunization with envelope subunit vaccine products

elicits neutralizing antibodies against laboratory-adapted but not primary isolates of human immunodeficiency virus type 1. The National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group', *J Infect Dis,* 173 (2), 340-8.

Mattapallil, J. J., et al. (2005), 'Massive infection and loss of memory CD4+ T cells in multiple tissues during acute SIV infection', *Nature,* 434 (7037), 1093-7.

Mayer, K. H., et al. (2003), 'Safety and tolerability of vaginal PRO 2000 gel in sexually active HIV-uninfected and abstinent HIV-infected women', *AIDS,* 17 (3), 321-9.

McCormack, S., et al. (2010), 'PRO2000 vaginal gel for prevention of HIV-1 infection (Microbicides Development Programme 301): a phase 3, randomised, double-blind, parallel-group trial', *Lancet,* 376 (9749), 1329-37.

McCutchan, F. E., et al. (2005), 'In-depth analysis of a heterosexually acquired human immunodeficiency virus type 1 superinfection: evolution, temporal fluctuation, and intercompartment dynamics from the seronegative window period through 30 months postinfection', *J Virol,* 79 (18), 11693-704.

McLean, A. R. and Nowak, M. A. (1992), 'Competition between zidovudine-sensitive and zidovudine-resistant strains of HIV', *AIDS,* 6 (1), 71-9.

McMichael, A. J. and Phillips, R. E. (1997), 'Escape of human immunodeficiency virus from immune control', *Annu Rev Immunol,* 15, 271-96.

Mehandru, S., et al. (2004), 'Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract', *J Exp Med,* 200 (6), 761-70.

Merz, B. (1987), 'HIV vaccine approved for clinical trials', *JAMA,* 258 (11), 1433-4.

Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990), 'DNA recombination during PCR', *Nucleic Acids Res,* 18 (7), 1687-91.

Miura, T., et al. (2009a), 'HLA-associated alterations in replication capacity of chimeric NL4-3 viruses carrying gag-protease from elite controllers of human immunodeficiency virus type 1', *J Virol,* 83 (1), 140-9.

--- (2008), 'Genetic characterization of human immunodeficiency virus type 1 in elite controllers: lack of gross genetic defects or common amino acid changes', *J Virol,* 82 (17), 8422-30.

--- (2009b), 'HLA-B57/B*5801 human immunodeficiency virus type 1 elite controllers select for rare gag variants associated with reduced viral replication capacity and strong cytotoxic T-lymphocyte [corrected] recognition', *J Virol,* 83 (6), 2743-55.

--- (2010), 'Impaired replication capacity of acute/early viruses in persons who become HIV controllers', *J Virol,* 84 (15), 7581-91.

Montefiori, D. C. (2005), 'Evaluating neutralizing antibodies against HIV, SIV, and SHIV in luciferase reporter gene assays', *Curr Protoc Immunol* (Chapter 12; Unit 12.11).

Montefiori, D. C., et al. (1996), 'Neutralizing and infection-enhancing antibody responses to human immunodeficiency virus type 1 in long-term nonprogressors', *J Infect Dis,* 173 (1), 60-7.

Moore, P. L., Gray, E. S., and Morris, L. (2009), 'Specificity of the autologous neutralizing antibody response', *Curr Opin HIV AIDS,* 4 (5), 358-63.

Moore, P. L., et al. (2008), 'The c3-v4 region is a major target of autologous neutralizing antibodies in human immunodeficiency

virus type 1 subtype C infection', *J Virol,* 82 (4), 1860-9.

Mselle, T. F., et al. (2009), 'Human uterine natural killer cells but not blood natural killer cells inhibit human immunodeficiency virus type 1 infection by secretion of CXCL12', *J Virol,* 83 (21), 11188-95.

--- (2007), 'Unique characteristics of NK cells throughout the human female reproductive tract', *Clin Immunol,* 124 (1), 69-76.

Murray, J. D. (2002), *Mathematical Biology I: An Introduction*, eds S. S. Antman, et al. (3rd edn., Interdisciplinary Applied Mathematics, 1; London: Springer) 551.

Naidu, Y. M., et al. (1988), 'Characterization of infectious molecular clones of simian immunodeficiency virus (SIVmac) and human immunodeficiency virus type 2: persistent infection of rhesus monkeys with molecularly cloned SIVmac', *J Virol,* 62 (12), 4691-6.

Nara, P. L., et al. (1987), 'Persistent infection of chimpanzees with human immunodeficiency virus: serological responses and properties of reisolated viruses', *J Virol,* 61 (10), 3173-80.

Nath, B. M., Schumann, K. E., and Boyer, J. D. (2000), 'The chimpanzee and other non-human-primate models in HIV-1 vaccine research', *Trends Microbiol,* 8 (9), 426-31.

National Institute of Allergy and Infectious Diseases, NIAID (2012), 'Statement: An update regarding the HVTN 502 and HVTN 503 HIV Vaccine Trials', (updated 22 October 2007) <http://www.niaid.nih.gov/news/newsreleases/2007/pages/step_update.aspx>, accessed 1st September 2012.

Nikolaitchik, O. A., et al. (2011), 'Multiple barriers to recombination between divergent HIV-1 variants revealed by a dual-marker recombination assay', *J Mol Biol,* 407 (4), 521-31.

Nishimura, Y., et al. (2005), 'Resting naive CD4+ T cells are massively

infected and eliminated by X4-tropic simian-human immunodeficiency viruses in macaques', *Proc Natl Acad Sci U S A,* 102 (22), 8000-5.

Novitsky, V., et al. (2011), 'Transmission of single and multiple viral variants in primary HIV-1 subtype C infection', *PLoS One,* 6 (2), e16714.

Nowak, M. A. and May, R. M. (1994), 'Superinfection and the evolution of parasite virulence', *Proc Biol Sci,* 255 (1342), 81-9.

O'Brien, S. J., Gao, X., and Carrington, M. (2001), 'HLA and AIDS: a cautionary tale', *Trends Mol Med,* 7 (9), 379-81.

O'Connell, K. A., et al. (2009), 'Role of natural killer cells in a cohort of elite suppressors: low frequency of the protective KIR3DS1 allele and limited inhibition of human immunodeficiency virus type 1 replication in vitro', *J Virol,* 83 (10), 5028-34.

Ochsenbauer, C., et al. (2012), 'Generation of transmitted/founder HIV-1 infectious molecular clones and characterization of their replication capacity in CD4 T lymphocytes and monocyte-derived macrophages', *J Virol,* 86 (5), 2715-28.

Otto, C., et al. (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 Day, T. (2007), A biologist's guide to mathematical modelling in ecology and evolution (Princeton University Press).

Pacold, M., et al. (2010), 'Comparison of methods to detect HIV dual infection', *AIDS Res Hum Retroviruses,* 26 (12), 1291-8.

Painter, S. L., et al. (2003), 'Envelope variants from women recently infected with clade A human immunodeficiency virus type 1 confer

distinct phenotypes that are discerned by competition and neutralization experiments', *J Virol,* 77 (15), 8448-61.

Palmer, S., et al. (2005), 'Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis', *J Clin Microbiol,* 43 (1), 406-13.

Pantaleo, G., et al. (1991), 'Dissociation between syncytia formation and HIV spreading. Suppression of syncytia formation does not necessarily reflect inhibition of HIV infection', *Eur J Immunol,* 21 (7), 1771-4.

Payne, R. P., et al. (2009), 'HLA-mediated control of HIV and HIV adaptation to HLA', *Adv Parasitol,* 68, 1-20.

Perelson, A. S., et al. (1996), 'HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time', *Science,* 271 (5255), 1582-6.

Pereyra, F., et al. (2008), 'Genetic and immunologic heterogeneity among persons who control HIV infection in the absence of therapy', *J Infect Dis,* 197 (4), 563-71.

Petit, S. J., Blondeau, C., and Towers, G. J. (2011), 'Analysis of the human immunodeficiency virus type 1 M group Vpu domains involved in antagonizing tetherin', *J Gen Virol,* 92 (Pt 12), 2937-48.

Phillips, A. N. (1996), 'Reduction of HIV concentration during acute infection: independence from a specific immune response', *Science,* 271 (5248), 497-9.

Phillips, R. E., et al. (1991), 'Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition', *Nature,* 354 (6353), 453-9.

Piantadosi, A., et al. (2009), 'HIV-1 evolution in gag and env is highly

correlated but exhibits different relationships with viral load and the immune response', *AIDS,* 23 (5), 579-87.

Pilcher, C. D., et al. (2004), 'Brief but efficient: acute HIV infection and the sexual transmission of HIV', *J Infect Dis,* 189 (10), 1785-92.

Pillay, D., et al. (2007), 'HIV phylogenetics', *BMJ,* 335 (7618), 460-1.

Pizzato, M., et al. (2009), 'A one-step SYBR Green I-based product-enhanced reverse transcriptase assay for the quantitation of retroviruses in cell culture supernatants', *J Virol Methods,* 156 (1-2), 1-7.

--- (2007), 'Dynamin 2 is required for the enhancement of HIV-1 infectivity by Nef', *Proc Natl Acad Sci U S A,* 104 (16), 6812-7.

Polacino, P., et al. (2008), 'Differential pathogenicity of SHIV infection in pig-tailed and rhesus macaques', *J Med Primatol,* 37 Suppl 2, 13-23.

Pope, M. and Haase, A. T. (2003), 'Transmission, acute HIV-1 infection and the quest for strategies to prevent infection', *Nat Med,* 9 (7), 847-52.

Prado, J. G., et al. (2009), 'Replicative Capacity of Mother-to-Child transmitted virus is associated with pediatric HIV-1 disease progression rate', *J Virol*.

Qureshi, H., et al. (2012), 'Low-dose penile SIVmac251 exposure of rhesus macaques infected with adenovirus type 5 (Ad5) and then immunized with a replication-defective Ad5-based SIV gag/pol/nef vaccine recapitulates the results of the phase IIb step trial of a similar HIV-1 vaccine', *J Virol,* 86 (4), 2239-50.

Rachinger, A., et al. (2008), 'Recovery of viremic control after superinfection with pathogenic HIV type 1 in a long-term elite controller of HIV type 1 infection', *Clin Infect Dis,* 47 (11), e86-9.

--- (2011), 'Time-measured phylogenies of gag, pol and env sequence data reveal the direction and time interval of HIV-1 transmission', *Aids,* 25 (8), 1035-9.

Rambaut, A. (2002), 'Se-Al. Sequence Alignment Editor v2.0a11'.

--- 'FigTree v1.3.1', <http://tree.bio.ed.ac.uk/software/figtree>.

Rambaut, A. and Drummond, A.J. 'Tracer v1.5', <http://tree.bio.ed.ac.uk/software/tracer>.

Rambaut, A., et al. (2004), 'The causes and consequences of HIV evolution', *Nat Rev Genet,* 5 (1), 52-61.

Ravet, S., et al. (2007), 'Distinctive NK-cell receptor repertoires sustain high-level constitutive NK-cell activation in HIV-exposed uninfected individuals', *Blood,* 109 (10), 4296-305.

Reeves, J. D., et al. (1999), 'Primary human immunodeficiency virus type 2 (HIV-2) isolates infect CD4-negative cells via CCR5 and CXCR4: comparison with HIV-1 and simian immunodeficiency virus and relevance to cell tropism in vivo', *J Virol,* 73 (9), 7795-804.

Rerks-Ngarm, S., et al. (2009), 'Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand', *N Engl J Med,* 361 (23), 2209-20.

Richman, D. D., et al. (2003), 'Rapid evolution of the neutralizing antibody response to HIV type 1 infection', *Proc Natl Acad Sci U S A,* 100 (7), 4144-9.

Rizzuto, C. D., et al. (1998), 'A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding', *Science,* 280 (5371), 1949-53.

Robey, W. G., et al. (1986), 'Prospect for prevention of human immunodeficiency virus infection: purified 120-kDa envelope glycoprotein induces neutralizing antibody', *Proc Natl Acad Sci U S*

A, 83 (18), 7023-7.

Roche (2005), Expand High Fidelity PCR System (Roche).

Rong, R., et al. (2009), 'Escape from autologous neutralizing antibodies in acute/early subtype C HIV-1 infection requires multiple pathways', *PLoS Pathog,* 5 (9), e1000594.

Ronquist, F. and Huelsenbeck, J. P. (2003), 'MrBayes 3: Bayesian phylogenetic inference under mixed models', *Bioinformatics,* 19 (12), 1572-4.

Ronquist, F., et al. (2012), 'MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space', *Syst Biol.*

Rowland-Jones, S., et al. (1995), 'HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women', *Nat Med,* 1 (1), 59-64.

Rybarczyk, B. J., et al. (2004), 'Correlation between env V1/V2 region diversification and neutralizing antibodies during primary infection by simian immunodeficiency virus sm in rhesus macaques', *J Virol,* 78 (7), 3561-71.

Salazar-Gonzalez, J. F., et al. (2008), 'Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing', *J Virol,* 82 (8), 3952-70.

--- (2009), 'Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection', *J Exp Med,* 206 (6), 1273-89.

Salemi, M., et al. (2007), 'Phylodynamics of HIV-1 in lymphoid and non-lymphoid tissues reveals a central role for the thymus in emergence of CXCR4-using quasispecies', *PLoS One,* 2 (9), e950.

--- (2009), 'Distinct patterns of HIV-1 evolution within metastatic

tissues in patients with non-Hodgkins lymphoma', *PLoS One,* 4 (12), e8153.

Salle, B., et al. (2010), 'Infection of macaques after vaginal exposure to cell-associated simian immunodeficiency virus', *J Infect Dis,* 202 (3), 337-44.

Santiago, M. L., et al. (2003a), 'Amplification of a complete simian immunodeficiency virus genome from fecal RNA of a wild chimpanzee', *J Virol,* 77 (3), 2233-42.

--- (2003b), 'Foci of endemic simian immunodeficiency virus infection in wild-living eastern chimpanzees (Pan troglodytes schweinfurthii)', *J Virol,* 77 (13), 7545-62.

Sauter, D., et al. (2009), 'Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains', *Cell Host Microbe,* 6 (5), 409-21.

Scarlatti, G., et al. (1996), 'Interplay of HIV-1 phenotype and neutralizing antibody response in pathogenesis of AIDS', *Immunol Lett,* 51 (1-2), 23-8.

Schmidt, H. A. and von Haeseler, A. (2007), 'Maximum-likelihood analysis using TREE-PUZZLE', *Curr Protoc Bioinformatics,* Chapter 6, Unit 6 6.

Schmidt, H. A., et al. (2002), 'TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing', *Bioinformatics,* 18 (3), 502-4.

Scott, G. B., et al. (1989), 'Survival in children with perinatally acquired human immunodeficiency virus type 1 infection', *N Engl J Med,* 321 (26), 1791-6.

Scott-Algara, D., et al. (2003), 'Cutting edge: increased NK cell activity in HIV-1-exposed but uninfected Vietnamese intravascular

drug users', *J Immunol,* 171 (11), 5663-7.

Shankarappa, R., et al. (1999), 'Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection', *J Virol,* 73 (12), 10489-502.

Sheehy, A. M., Gaddis, N. C., and Malim, M. H. (2003), 'The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif', *Nat Med,* 9 (11), 1404-7.

Sheehy, A. M., et al. (2002), 'Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein', *Nature,* 418 (6898), 646-50.

Simmonds, P., et al. (1990), 'Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1', *J Virol,* 64 (12), 5840-50.

Smith, D. M., Richman, D. D., and Little, S. J. (2005a), 'HIV superinfection', *J Infect Dis,* 192 (3), 438-44.

Smith, D. M., et al. (2005b), 'HIV drug resistance acquired through superinfection', *AIDS,* 19 (12), 1251-6.

Smith, J. M. and Haigh, J. (1974), 'The hitch-hiking effect of a favourable gene', *Genet Res,* 23 (1), 23-35.

Spira, A. I., et al. (1996), 'Cellular targets of infection and route of viral dissemination after an intravaginal inoculation of simian immunodeficiency virus into rhesus macaques', *J Exp Med,* 183 (1), 215-25.

Stafford, M. A., et al. (2000), 'Modeling Plasma Virus Concentration during Primary HIV Infection', *J Theor Biol,* 203 (3), 285-301.

Stephenson, K. E., et al. (2012), 'Preexisting adenovirus seropositivity is not associated with increased HIV-1 acquisition in three HIV-1 vaccine efficacy trials', *J Infect Dis,* 205 (12), 1806-10.

Stevens, C. E., et al. (1980), 'Hepatitis B vaccine: immune responses in haemodialysis patients', *Lancet,* 2 (8206), 1211-3.

Stone, M., et al. (2010), 'A limited number of simian immunodeficiency virus (SIV) env variants are transmitted to rhesus macaques vaginally inoculated with SIVmac251', *J Virol,* 84 (14), 7083-95.

Streeck, H., et al. (2008), 'Immune-driven recombination and loss of control after HIV superinfection', *J Exp Med,* 205 (8), 1789-96.

--- (2009), 'Human immunodeficiency virus type 1-specific CD8+ T-cell responses during primary infection are major determinants of the viral set point and loss of CD4+ T cells', *J Virol,* 83 (15), 7641-8.

Strickland, S. L., et al. (2011), 'Significant genetic heterogeneity of the SIVmac251 viral swarm derived from different sources', *AIDS Res Hum Retroviruses,* 27 (12), 1327-32.

Strimmer, K. and von Haeseler, A. (1997), 'Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment', *Proc Natl Acad Sci U S A,* 94 (13), 6815-9.

Svara, F. and Rankin, D. J. (2011), 'The evolution of plasmid-carried antibiotic resistance', *BMC Evol Biol,* 11 (1), 130.

Szmuness, W., et al. (1980), 'Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States', *N Engl J Med,* 303 (15), 833-41.

Tamura, K. and Nei, M. (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees', *Mol Biol Evol,* 10 (3), 512-26.

Thali, M., et al. (1991), 'Characterization of a discontinuous human immunodeficiency virus type 1 gp120 epitope recognized by a broadly reactive neutralizing human monoclonal antibody', *J Virol,*

65 (11), 6188-93.

Thippeshappa, R., et al. (2011), 'Vif substitution enables persistent infection of pig-tailed macaques by human immunodeficiency virus type 1', *J Virol,* 85 (8), 3767-79.

Troyer, R. M., et al. (2005), 'Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression', *J Virol,* 79 (14), 9006-18.

UNAIDS (2010), 'Global Report: UNAIDS Report on the Global AIDS Epidemic 2010', (Geneva: United Nations).

Valentin, A. and Pavlakis, G. N. (2003), 'Natural killer cells are persistently infected and resistant to direct killing by HIV-1', *Anticancer Res,* 23 (3A), 2071-5.

Valenzuela, A., et al. (1997), 'Neutralizing antibodies against the V3 loop of human immunodeficiency virus type 1 gp120 block the CD4-dependent and -independent binding of virus to cells', *J Virol,* 71 (11), 8289-98.

Van Damme, L., et al. (2002), 'Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomised controlled trial', *Lancet,* 360 (9338), 971-7.

van der Kuyl, A. C. and Cornelissen, M. (2007), 'Identifying HIV-1 dual infections', *Retrovirology,* 4, 67.

van der Kuyl, A. C., et al. (2010), 'Analysis of infectious virus clones from two HIV-1 superinfection cases suggests that the primary strains have lower fitness', *Retrovirology,* 7, 60.

Volberding, P. A. and Deeks, S. G. (2010), 'Antiretroviral therapy and management of HIV infection', *Lancet,* 376 (9734), 49-62.

Wadman, M. (2011), 'Animal rights: Chimpanzee research on trial', *Nature,* 474 (7351), 268-71.

Waterhouse, A. M., et al. (2009), 'Jalview Version 2--a multiple sequence alignment editor and analysis workbench', *Bioinformatics,* 25 (9), 1189-91.

Wawer, M. J., et al. (2005), 'Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda', *J Infect Dis,* 191 (9), 1403-9.

Wei, B. L., et al. (2005), 'Inhibition of lysosome and proteasome function enhances human immunodeficiency virus type 1 infection', *J Virol,* 79 (9), 5705-12.

Wei, X., et al. (2003), 'Antibody neutralization and escape by HIV-1', *Nature,* 422 (6929), 307-12.

West, S. A., El Mouden, C., and Gardner, A. (2011), 'Sixteen common misconceptions about the evolution of cooperation in humans', *Evol Hum Behav,* 32 (4), 231-62.

Willey, R. L., et al. (1986), 'Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome retrovirus', *Proc Natl Acad Sci U S A,* 83 (14), 5038-42.

Williams, B. G., et al. (2011), 'Epidemiological impact of tenofovir gel on the HIV epidemic in South Africa', *J Acquir Immune Defic Syndr,* 58 (2), 207-10.

Wilson, D. J. and McVean, G. (2006), 'Estimating diversifying selection and functional constraint in the presence of recombination', *Genetics,* 172 (3), 1411-25.

Wilson, D. P., et al. (2008), 'Relation between HIV viral load and infectiousness: a model-based analysis', *Lancet,* 372 (9635), 314-20.

Wolinsky, S. M., et al. (1992), 'Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants',

Science, 255 (5048), 1134-7.

World Health Organization, WHO (1st September 2012), 'HIV/AIDS: Condoms for HIV prevention', (updated 25th March 2009) <http://www.who.int/hiv/topics/condoms/en/>.

Wright, S. (1938), 'Size of population and breeding structure in relation to evolution', *Science,* 87, 430-31.

--- (1948), 'On the roles of directed and random changes in gene frequency in the genetics of populations', *Evolution,* 2 (4), 279-94.

Yang, G. B., et al. (2012), 'Immunization with recombinant macaque major histocompatibility complex class I and II and human immunodeficiency virus gp140 inhibits simian-human immunodeficiency virus infection in macaques', *J Gen Virol,* 93 (Pt 7), 1506-18.

Yang, S. J., et al. (2011), 'Lack of adaptation to human tetherin in HIV-1 group O and P', *Retrovirology,* 8, 78.

Yang, Z., Goldman, N., and Friday, A. (1995), 'Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem', *Systematic Biology,* 44 (3), 384-99.

Yawata, M., et al. (2006), 'Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function', *J Exp Med,* 203 (3), 633-45.

Yeh, W. W., et al. (2010), 'Autologous neutralizing antibodies to the transmitted/founder viruses emerge late after SIVmac251 infection of rhesus monkeys', *J Virol.*

Zhang, F., et al. (2008), 'No effect of endogenous TRIM5alpha on HIV-1 production', *Nat Med,* 14 (3), 235-6; author reply 36-8.

Zhang, J. and Temin, H. M. (1994), 'Retrovirus recombination depends on the length of sequence identity and is not error prone', *J*

Virol, 68 (4), 2409-14.

Zhang, L. Q., et al. (1993), 'Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection', *J Virol,* 67 (6), 3345-56.

Zhang, M., et al. (2004), 'Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin', *Glycobiology,* 14 (12), 1229-46.

Zhu, P., et al. (2006), 'Distribution and three-dimensional structure of AIDS virus envelope spikes', *Nature,* 441 (7095), 847-52.

Zhu, T., et al. (1993), 'Genotypic and phenotypic characterization of HIV-1 patients with primary infection', *Science,* 261 (5125), 1179-81.

--- (1996), 'Genetic characterization of human immunodeficiency virus type 1 in blood and genital secretions: evidence for viral compartmentalization and selection during sexual transmission', *J Virol,* 70 (5), 3098-107.

Zuckerkandl, E. and Pauling, L. (1962), 'Molecular disease, evolution, and genic heterogeneity', in M.A. Kasha and B. Pullman (eds.), *Horizons in Biochemistry: Albert Szent-Györgyi dedicatory volume* (London: Academic Press), 189-225.

Zuckerman, R. A., et al. (2004), 'Higher concentration of HIV RNA in rectal mucosa secretions than in blood and seminal plasma, among men who have sex with men, independent of antiretroviral therapy', *J Infect Dis,* 190 (1), 156-61.