**Detecting recombination and its mechanistic association with genomic features via statistical models**

by

Misha Rajaram

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Karin S. Dorman, Co-major Professor
Dennis Lavrov, Co-major Professor
Bryony Bonning
Xun Gu
Peng Liu

Iowa State University

Ames, Iowa

2010

# DEDICATION

*For Appa and Amma*

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGEMENTS

The journey leading up to the milestone that is this thesis has been a deeply satisfying experience. In all these years many people have been responsible for directly and indirectly supporting and influencing my academic career. To all those people, I would like to express my heartfelt gratitude.

First I would like to thank my adviser Dr. Karin Dorman for her indispensable contributions towards the research presented in the following pages. For always leading by example, for her strong work ethic and seemingly unlimited patience, I am very grateful. I am also thankful to her for believing in me and for all her timely words of encouragement and support.

I would also like to thank members of my committee Dr. Dennis Lavrov, Dr. Bryony Bonning, Dr. Xun Gu and Dr. Peng Liu for their guidance and lively discussions at committee meetings. My special thanks are due to Dr. Susan Carpenter and Dr. Drena Dobbs for providing me with valuable insights into the biology of retrovirus evolution.

I would like to thank Trish Stauble of the Bioinformatics and Computational Biology program for being ever willing to help with every problem, small or big, and for her expert guidance with all the paperwork. To all my friends from Ames, in particular, Grishma, Abhinav, Xiao, Vishwa, Vani, Abhijeet, Prem, Saikat, Sumohan, Arpita, Kritanjali, Bandy, Ritu, Benazir, Firoz, Srini and Savitha, I would like to express my gratitude for making these past five years most memorable and for standing by me through them. My thanks are also due to Drs. Krishna and Rani Athreya for their love and for helping me keep in touch with music through the Raga group. To Drs. Rema and Sree Nilakanta

and Haema I owe a deep debt of gratitude for opening their home and hearts to me and treating me like family.

Finally, I would like to thank my family for their unconditional love and support of all that I have endeavored. To Thatha and Paati I am thankful for always magnifying my sense of achievement with the pride they feel. To my sister Megha, I am thankful for always being a willing ally in my plans for world domination. I am grateful to my husband Ankur for his patient support and encouragement through my many moments of self-doubt. Finally, I am thankful to my parents for everything they have done for me, for the many sacrifices they have had to make and for always believing in me.

# ABSTRACT

Recombination is a powerful weapon in the evolutionary arsenal of retroviruses such as HIV. It enables the production of chimeric variants or recombinants that may confer a selective advantage to the pathogen over the host immune response. Recombinants further accentuate differences in virulence, disease progression and drug resistance mutation patterns already observed in non-recombinant variants of HIV. This thesis describes the development of a rapid genotyper for HIV sequences employing supervised learning algorithms and its application to complex HIV recombinant data, the application of a hierarchical model for detection of recombination hotspots in the HIV-1 genome and the extension of this model enabling estimation of the association between recombination probabilities and covariates of interest.

The rapid genotyper for HIV-1 explores a solution to the genotyping problem in the machine learning paradigm. Of the algorithms tested, the genotyper built using Bayesian additive regression trees (BART) was most successful in efficiently classifying complex recombinants that pose a challenge to other currently available genotyping methods. We also developed a novel method, bootSMOTE, for generating synthetic data in order to supplement insufficient training data. We found that supplementation with synthetic recombinants especially boosts identification of complex recombinants. We describe the genotyper software available for download as well as a web interface enabling rapid classification of HIV-1 sequences.

Hotspots for recombination in the HIV-1 genome are modeled using spatially smoothed changepoint processes. This hierarchical model uses a phylogenetic recombination de-

tection model of dual changepoint processes at the lower level. The upper level applies a Gaussian Markov random field (GMRF) hyperprior to population-level recombination probabilities in order to efficiently combine the information from many individual recombination events as inferred at the lower level. Focusing on 544 unique recombinant sequences, we found a novel hotspot in the *pol* gene of HIV-1 while confirming the presence of high recombination activity in the *env* gene.

Valuable insights into the molecular mechanism of recombination may be gained by extending the GMRF model to include covariates of interest. We add a level to the hierarchical model and allow for the simultaneous inference of recombination probabilities as well their association with genomic covariates of interest. Using a set of 527 unique recombinants, we confirmed the presence of the *pol* hotspot. Interestingly, we found significant positive associations of spatial fluctuations in recombination probabilities with genomic regions prone to forming secondary structure as well as significant negative associations with regions that support tight RNA-DNA hybrid formation. Overall, our results support the theory that pause sites along the genome promote recombination.

# CHAPTER 1. GENERAL INTRODUCTION

## 1.1 Introduction

The persistence of the Human Immunodeficiency Virus (HIV) as a major pathogen causing a global pandemic bears testimony to the success of the molecular mechanisms promoting its rapid evolution (Burke, 1997). HIV is a *lentivirus* that can stay apparently dormant in the host for many years after infection (Vogt, 1997). Eventually, it leads to the Acquired Immunodeficiency Syndrome, characterized by an attenuated host immune system. The infected individual becomes increasingly susceptible to a multitude of opportunistic infections that prove fatal (Weiss, 1993; Douek et al., 2009).

HIV packages its genetic material in the form of two single stranded positive sense RNA molecules deposited in the host cell upon infection (Coffin, 1979). The reverse transcription stage in its replication cycle is responsible for creating a double stranded DNA copy using a packaged RNA molecule as template. Integration of viral DNA with the host genome enables it to use host resources for production of HIV proteins (Temin, 1991). The reverse transcription machinery introduces $3\times10^{-5}$ mutations per nucleotide per cycle of replication (Levy et al., 2004) in the $10^8-10^9$ virions produced per day. Such high rates of replication and mutation together with the recombinogenic nature of the replication complex contribute to the rich genetic diversity of the virus (Robertson et al., 1995). Recombination in HIV refers to a mechanism wherein the reverse transcription complex is prone to switching between the two viral RNA templates present in the host cell. This leads to chimeric or recombinant DNA molecules that are mosaics of the two

RNA molecules packaged by the infecting virion (Negroni and Buc, 2001). Inter-subtype recombination occurs when the two RNA templates used are genetically distinct. As of 2010, nearly 50 stable genetically distinct variants or genotypes of HIV-1 have been identified (HIV-Database, 2010). Of these, 11 are non-recombinant and referred to as subtypes while the rest are inter-subtype recombinants known as Circulating Recombinant Forms (CRFs) (McCutchan, 2006). Also present are "transient" recombinants or Unique Recombinant Forms (URFs) believed to be products of isolated recombination events. Genotypes display marked differences in virulence, disease progression, preferred mode of transmission, drug resistance mutation profiles as well as sensitivity to detection assays (Spira et al., 2003; Madani et al., 2004; Baeten et al., 2006; Swanson et al., 2005; Colson et al., 2007).

Identifying a recombinant, its parental genotypes and the location of the recombination breakpoints, is therefore, not only essential from the perspective of efficient clinical management (Geretti, 2006) but is also imperative towards better understanding of the recombination mechanism. Galetto et al. (2006) point out that in order to understand the mechanism of recombination, it is important to ascertain causes or triggers that promote template switching by the replication complex. Further, experimental and computational analyses (Dykes et al., 2004; Balakrishnan et al., 2003; Magiorkinis et al., 2003; Fan et al., 2007; Galli. et al., 2008) have shown that the location of recombination breakpoints is not random along the genome. This suggests that some regions of the genome, by virtue of sequence, structural or functional features, favor recombination.

Many genotyping techniques exist (McGuire et al., 1997; Husmeier and Wright, 2001; Rozanov et al., 2004; de Oliveira et al., 2005; Wu et al., 2007). Most are distance or phylogeny based and use a sliding window technique (Grassly and Holmes, 1997; Husmeier and Wright, 2001; Husmeier and McGuire, 2003). In most cases, the strategy is to use a reference set of known genotypes and compare a query sequence with it using some measure of relatedness. Such methods are able to predict the occurrence

of recombination and the parental genotypes. However, in recent years there has been a rise in the number of recombinant sequences (HIV-Database, 2010). Most existing genotyping methods find the task of genotyping complex recombinant sequences very challenging (Holguin et al., 2008a). As the viral genome gets more and more complex, simple relatedness matrices prove insufficient in providing information useful for accurate genotyping. A more sophisticated approach is to use these matrices to look for patterns that might appear to be peculiar to a genotype. Supervised learning algorithms provide the apt paradigm for such exploration. The first part of this thesis presents a novel application of supervised learning algorithms to rapid genotyping of HIV-1 sequences.

While genotypers can identify parental genotypes, they lack the ability to accurately identify the position of recombination breakpoints in the genome. Suchard et al. (2002) provided a Bayesian framework for inference of recombination breakpoint locations. Minin et al. (2005) and Fang et al. (2007) developed this model further by de-convoluting other nuisance parameters from the parameters of interest: the number and locations of recombination breakpoints. Being able to accurately estimate the recombination breakpoints puts us in the advantageous position of being able to combine this information from many sequences and gain insight into the spatial variation of recombination probabilities in the HIV genome. To achieve this, Minin et al. (2007) describe a hierarchical model that places a spatially smoothing Gaussian Markov Random Field (GMRF) prior on the recombination probabilities in order to effectively combine information from analyses of individual sequences. We developed this model further to enhance its ability to deal with more diverse datasets. Further, we extended the model to be able to simultaneously infer the association of the recombination probabilities with covariates. This extension provides, for the first time, a unified framework to establish a relationship between recombination probabilities and genomic features that may be associated with recombination.

## 1.2 Thesis Organization

This thesis is presented as a series of manuscripts and is organized as follows. Background material is presented in two review chapters. Chapter 2 reviews HIV biology and machine learning algorithms while Chapter 5 reviews statistical techniques used in the development of models in later chapters.

Chapter 3 describes the application of supervised learning algorithms to the problem of rapidly genotyping HIV-1 sequences. We build a classifier using a setup of parallel binary classifiers, one for each genotype in the HIV reference set. Relatedness of a sequence to the reference sequences is summarized using various metrics producing feature sets that form suitable input for the binary classifiers. Next we describe a novel method to produce synthetic sequences suitable for data supplementation, bootSMOTE. Finally, the chapter provides details of the genotyper software and its web interface. This manuscript was prepared by authors Misha L. Rajaram (MLR) and Karin S. Dorman (KSD) who also jointly conceptualized and designed the study. MLR implemented the classifier and performed the experiments. Yves Sucaet implemented the web interface.

Chapter 4 presents methods for careful curation of large HIV datasets. It also shows results from the application of the model described by Minin et al. (2007) to large datasets. Authors Vladimir N. Minin (VNM), Marc A. Suchard and KSD were responsible for conceptualizing the study. MLR and KSD designed the study and prepared the manuscript. MLR implemented the clustering algorithm and obtained results presented herein.

Chapter 6 presents the extension of the hierarchical model to include covariate inference. We present results of application of the model to a dataset of HIV sequences to infer the spatial preference for recombination probabilities as well as their association with some genomic covariates. Authors KSD and VNM initially conceived of the extension model, which was later greatly adapted by KSD and MLR. MLR and KSD de-

signed the experiments and prepared the manuscript. MLR implemented the model and conducted experiments. Susan Carpenter and Drena Dobbs were involved in long discussions about this work, and in particular suggested the secondary structure covariates to include in the model and guided interpretation of results.

Chapter 7 presents an alternate hierarchical GMRF model specially suited to deal with binomial responses. This model doubles the speed of inference while providing variance stabilized data. Finally, Chapter 8 presents general conclusions and plans for future work.

# CHAPTER 2. REVIEW I

## 2.1 The Human Immunodeficiency Virus (HIV)

Group : **Group IV (ssRNA-RT)**

Family: **Retroviridae**

Subfamily: **Orthoretrovirinae**

Genus: **Lentivirus**

Species: **Human Immunodeficiency Virus**

The Human Immunodeficiency Virus (HIV), is a *lentivirus* belonging to family *Retroviridae*. Retroviruses are RNA viruses that replicate via the *reverse transcriptase* enzyme when it creates DNA copies from their RNA genome. Enzyme *integrase* then facilitates the integration of viralDNAwith the host genome. Thereafter, the viral DNA uses the host machinery for replication. Lentiviruses are characterized by long incubation periods wherein the integrated viral DNA lies clinically dormant (Desport, 2010).

HIV infects cells vital to the human immune response, namely, CD4+ T helper cells, macrophages and dendritic cells. The infection progresses in three main stages. The first or primary infection stage, occurs $2-4$ weeks post-exposure and is marked by a sharp drop in numbers of circulating CD4+ cells. Most infected individuals at this stage develop influenza-like viremia caused upon activation of CD8+ T cells that kill infected cells (Kahn and Walker, 1998). The second stage is asymptomatic and can last between two weeks and 20 years. At this stage, the virus is found predominantly in immune cells and lymph nodes. Infected individuals, while capable of passing on the infection,

display no clinical symptoms. The third stage is characterized by a rapid decline in CD4+ cell count leading to the emergence of symptoms related to mild opportunistic infections. The final stage, known as the Acquired Immunodeficiency Syndrome (AIDS) results from loss of cell mediated immunity due to very low CD4+ cell counts ($\leq$ 200 cells/$\mu$l). Such extensive damage to the host immune system results in increasingly severe opportunistic infections that eventually prove fatal (WHO, 2007).

As of 2006, 0.6% of the world's population was infected with HIV and since its discovery the infection is estimated to have claimed 25 million lives (UNAIDS, 2006). HIV is transmitted through infected blood, semen, vaginal fluid and breast milk. The main modes of transmission include unsafe sex with infected individuals, use of contaminated needles, through breast milk and from an infected mother to her child at birth (Kilmarx, 2008). Antiretroviral therapy serves to delay the onset of AIDS in infected individuals and is known to prevent the transmission from mother to child. However, the virus is also accumulating resistance mutations to these drugs often rendering them ineffective.

### 2.1.1 Structure and Genome Organization of HIV

The HIV virion is spherical and $\sim$ 120 nm in diameter. At the core the *p24* viral proteins form a conical capsid. The capsid encloses two positive sense single stranded RNA molecules found tightly bound with nucleocapsid proteins *p6* and *p7* as well as the *reverse transcriptase* and *integrase* enzymes. The nucleocapsid proteins serve to protect the RNA from digestion by nucleases. Also present in the capsid are other enzymes and proteins such as *protease, vif, rev* and *nef* required at various stages of viral replication. Surrounding the capsid is a matrix formed of *p17* protein particles. The matrix is responsible for the integrity of the virion. Covering the matrix, is the envelope, formed when the capsid buds out of the host cell. The function of the envelope is to ensure attachment to a host cell. Towards this, the envelope has several spike-like structures made of viral glycoprotein particles, *gp120* and *gp41*. *gp120* or SU is present on the

surface and *gp41* or TM , a trans-membrane glycoprotein serves to anchor the surface protein and are in turn held in place by the matrix proteins (Kuiken et al., 2010).

Fig. 2.1 shows the linear genome organization of the virus. The HIV genome is $\sim$ 9.7 kb long and comprised of three main retroviral genes *gag*, *pol* and *env* which contain all the structural information required to produce new viral particles. The *gag* (group specific antigen) gene codes for the capsid and matrix proteins. The gene product is a fusion protein *p55* that is later cleaved by the *protease* enzyme to form *p17, p24, p6* and *p7*. The *pol* gene codes for the enzymes required in the viral life cycle. The *reverse transcriptase* (RT) enzyme, an RNA dependent DNA polymerase, is responsible for producing double stranded cDNA copies of the viral RNA. *Integrase* is required for integration of the viral DNA with the host DNA and *protease* is required for processing viral gene products to make them suitable for packaging and assembly of new virions. The *gag* and *pol* genes are adjacent on the genome and include a small overlap region. The 55 kDa *gag* precursor protein is produced by conventional translation. At a rate 20-fold lower, a programmed -1 frameshift during translation produces the *gag-pol* polyprotein. The polyprotein is then cleaved to form the assembly proteins and the enzymes (Hill et al., 2005).

The third major gene, *env* codes for a glycoprotein *gp160* that gives rise, upon cleavage, to the envelope proteins *gp120* and *gp41*. There are six other genes in the HIV genome that have regulatory or accessory functions. The *tat* or transactivator gene is responsible for promoting transcription of the virus. The first 59 nucleotides of the HIV genome contain the transactivator activation region or "TAR". Binding of the *tat* gene product to the TAR element serves to accelerate transcription of the viral genome integrated with the host DNA. The *rev* gene is responsible for nuclear export of viral mRNA. It binds to the *rev responsive element* or RRE on the intact mRNA and transports it out of the nucleus. *Nef* or negative replication factor is a negative regulator of replication. *Vif* or viral infectivity factor is believed to increase infectivity

of the virion. Its gene product binds to host APOBEC3G that is aimed at rendering the virion non-productive. The *vpr* or viral protein R is responsible for nuclear localization of the pre-integration complex that consists of the viral RNA, reverse transcriptase and the integrase enzymes. The final regulatory gene is *vpu* in HIV-1 and *vpx* in HIV-2. *vpu* is believed to be involved in assembly and packaging of new viral particles while *vpx* seems to have functions similar to *vpr* (Kuiken et al., 2010).

### 2.1.2 HIV Replication

Fig. 2.2 shows a simplified version of the life cycle of the virus in a host cells. The infective cycle of the virus starts with attachment of the virion via the gp120 surface glycoprotein to host cells presenting the CD4+ antigen receptors on their surface. This mainly includes two types of human immune cells– CD4+ T lymphocytes and macrophages. Chemokine receptors present on host cells, CCR5 and CXCR4, act as co-receptors that strengthen the attachment of the virion (Chan and Kim, 1998; Wyatt and Sodroski, 1998). Attachment prompts fusion of the viral envelope to the cell membrane of the host cell, releasing the contents of the viral capsid into the host cytoplasm (Coffin et al., 1997).

The next step is the reverse transcription of the RNA genome by *reverse transcriptase*. The replication machinery also degrades the RNA template as it moves along, making the DNA copy. Next a complementary DNA strand is added to form a double stranded DNA copy of the viral genome. This double stranded DNA is then inserted in the host DNA with the help of the *integrase* enzyme. At this stage this integrated viral genome is referred to as the *provirus*. Upon suitable activation, the host replication machinery starts making RNA copies of the viral DNA. These are transported to the cytoplasm and corresponding protein products produced. The *protease* enzyme cleaves these products to produce proteins that are then packaged and assembled into new virions. Virions are released from the host cell through budding (Coffin et al., 1997).

## 2.2 Genetic Diversity of HIV

The combination of a fast replication cycle and lack of a proof reading mechanism in the replication machinery affords the virus a great evolutionary advantage as evidenced by its rich genetic diversity. The replication cycle of the virus is very fast, producing between $10^8$ and $10^9$ virions per day (Robertson et al., 1995; Rambaut et al., 2004). The reverse transcription produces replicates with a high mutation rates of over $3 \times 10^{-5}$ per nucleotide per replication cycle (Levy et al., 2004). HIV has two main types: HIV-1 that is responsible for the current pandemic and HIV-2 that is known to cause infections restricted to parts of West Africa. Compared to HIV-1, HIV-2 shows slower disease progression and is often not as infective as HIV-1 (MacNeil et al., 2007).

HIV-1 is further divided into three groups: M (Major) causing over 90% of all HIV infections, N (Non-major) and O (Other) that are also found isolated in Africa (Robertson et al., 2000). In 2009, an isolate closely related to gorilla Simian Immunodeficiency virus was found in a Cameroonian woman and designated as HIV -1 Group P (Plantier et al., 2009). Further, group M has 11 genetically distinct subtypes, A-D, F-H, J-K with subtypes A and F divided into subtypes A1/A2 and F1/F2 (Robertson et al., 2000). The inter-subtype genetic diversity can be as high as 30% in the *env* gene and about 15% in the *gag* gene (Spira et al., 2003). Fig. 2.3 shows the global distribution of HIV subtypes. Note that they tend to be geographically isolated except in Africa which has a substantial representation of all subtypes.

### 2.2.1 Recombination in HIV

The genetic diversity of HIV is further augmented by the recombinogenic properties of the replication machinery. Co-infected or super-infected host cells are prone to produce virions that co-package genetically distinct genotypes (Smith et al., 2005; Chohan et al., 2005; Piantadosi et al., 2007). Upon infection by such a heterozygous virion, there is a

high likelihood of the production of recombinants during reverse transcription.

**Mechanism of recombination:** Several competing hypotheses exist on the mechanism of recombination. Of these, the most popular is the "copy choice" model (Coffin, 1979) that postulates that recombination occurs when the reverse transcription machinery switches templates midway during the production of the minus strand cDNA resulting in a mosaic genome that has fragments from both RNA molecules co-packaged by the infecting virion. Fig 2.4 illustrates the copy choice model for recombination. Further, most recombination events seem to occur during minus strand synthesis (Anderson et al., 1998; Zhang et al., 2000). Variations to the "copy choice" model serve to incorporate factors that influence various stages of the process such as the reverse transcriptase activity (Chen et al., 2003; DeStefano et al., 1996; Diaz and DeStefano, 1996; Roda et al., 2002a; Wu et al., 1995), secondary structure formation (Andersen et al., 2003; Balakrishnan et al., 2001, 2003; Moumen et al., 2001; Negroni and Buc, 2000) and viral nucleocapsid proteins (Negroni and Buc, 1999; Roda et al., 2003).

**Pausing and recombination rates:** The presence of pause sites along the genome has been implicated as one of the major factors affecting recombination rates. Homopolymeric G and C stretches as well as secondary structures have been shown to cause pausing of the reverse transcription machinery (Klarmann et al., 1993; Suo and Johnson, 1997; Wu et al., 1995). *In vitro* studies show that such pause sites are positively associated with strand transfer events (Roda et al., 2002a, 2003; Zhuang et al., 2002). Lanciault and Champoux (2006) reported that unpaired nucleotides were associated with lowered recombination rates *in vitro*. Similar conclusions were made by Dykes et al. (2004) for crossover points in the *gag* gene.

**Types of recombinants:** The spread of the AIDS pandemic has also seen the rise, in recent years, of recombinant forms that have spread through a population and caused

local epidemics. Such inter-subtype recombinants are termed as **Circulating Recombinant Forms** (CRFs) and currently, there are 47 identified types (HIV-Database, 2010). A recombinant is termed a CRF if three or more epidemiologically unrelated individuals are found infected by it. All other inter-subtype recombinants are referred to as **Unique Recombinant Forms** (URFs).

### 2.2.2 Implications of Genetic Diversity

Genetic diversity of the virus in the form of subtypes as well as emerging recombinant forms confers the virus with a great selective advantage over the host immune response. Local epidemics caused by recombinants are found to be spreading (Tovanabutra et al., 2004; Ramirez et al., 2008) with some evidence of recombinant infections taking over pure subtype infections (Spira et al., 2003). Buonaguro et al. (2007) showed that about 18% of infections worldwide were caused by recombinant forms. More recently, new URFs subtypes B, C and D have been characterized in north India (Bano et al., 2009), and there is evidence of widespread infection by CRF35_AD in Afghanistan (Sanders-Buell et al., 2010). As many as 10 new CRFs have been identified in the past two years (Kuiken et al., 2010). The genetic diversity in the virus manifests as phenotypic diversity in mode of transmission, co-receptor usage, disease progression, virulence, replicative fitness and drug resistance (Spira et al., 2003).

**Cell tropism and Co-receptor preference:** CCR5 and CXCR4 are chemokine receptors present on the host cell surface that act as co-receptors to the infecting virion. Viruses that prefer the CCR5 co-receptor are designated R5 viruses, those that prefer CXCR4 co-receptor are grouped as X4 viruses and those that use both are called R5X4 viruses (Clapham and McKnight, 2001). Early *in vitro* experiments performed using blood lymphocytes and T-cell lines confirmed that some isolates preferred to infect macrophages (M-tropic) while others preferred CD4+ T cells as hosts (T-tropic). A

third kind that could infect both was designated to be dual or D-tropic (Bagnarelli et al., 2004). These experiments, along with the discovery of a high correlation between CXCR4 use and T-tropism lead to the long-held belief that co-receptor usage is explained completely by cell tropism with M-tropic viruses showing R5 preference and T-tropic isolates showing X4 preference (Collman et al., 1989; Gendelman et al., 1988).

However, many studies have found that *in vivo*, T-cells and macrophages express both types of co-receptors on their surface. While the majority preference for primary infection is the CCR5 co-receptor, CD4+ T cells expressing CCR5 may also be infected by R5 viruses. Further, studies have also shown that X4 viruses may just as efficiently use macrophage X4 for infection, making them T-tropic or D-tropic (reviewed in Goodenow and Collman (2006)).

Most studies involving co-receptor usage have been performed using subtypes B and C. A majority of subtype B isolates start out as R5 populations. After years of chronic infection, about 50% subtype B isolates are believed to favor CXCR4 usage (Bratt et al., 1998; Peeters et al., 1999) while this proportion is lower (0-30%) in subtype C isolates (Ping et al., 1999). Dominant X4 populations were found in CRF_01AE samples (Tscherning et al., 1998; Yu et al., 1995) and another found X4 use more common during primary infection with subtype D than subtype A (Kaleebu et al., 2002). Esbjornsson et al. (2010) also found high X4 preference in CRF_02AG isolates.

R5 viruses are associated with a more stable infection with co-receptor switch occurring in later stages of the disease. X4 viruses on the other hand are associated with rapid CD4+ T cell count decline and disease progression (Philpott, 2003). Interestingly, association of mode of transmission with co-receptor usage has also been documented with R5 viruses believed to be predominant during sexual trasmission (Zhu et al., 1996)

**Mode of transmission:** While differences have been found in preference of mode of transmission by specific genotypes, much contradictory knowledge exists on the sub-

ject. Studies showed that in pregnant women in Kenya infected with subtypes A, C and D, infections from subtype C were found to be more advanced than A and D (Soto-Ramirez et al., 1996; Baeten et al., 2006). Another Kenyan study showed that subtype C had 6.1 times higher transmission potential from infected mother to unborn child than subtype D (Neilson et al., 1999). Later studies on the same population, however, found no significant difference (Blackard et al., 2001). Incidence of multiple infections and hence recombinants has been found to be high in Injection Drug Users (Templeton et al., 2009).

**Testing and Diagnosis:** Most tests developed for detection of the virus in a host system were developed in the western world and tailor-made to identify subtype B infections accurately. As early as the 1990s, HIV-1 group O tested negative with assays using subtype B epitopes (Gaschen et al., 2002). Much progress has been made with many second and third generations tests that are able to identify all subtypes as well as identify early phase and late phase antibodies (Iweala, 2004). However, these do not have 100% sensitivity and results vary in the detection of non-B subtypes (Holguin et al., 2008a).

**Anti-retroviral drug therapy:** Current drug therapy is aimed at interfering with the viral replication cycle at various stages. Reverse transcriptase inhibitors and Protease inhibitors are among the most common therapies and were, like diagnostic tests, developed in the western world. In patients with and without subtype B infections, the effect of these drugs and the development of drug resistant mutations is a cause for concern. It is interesting to note that even before the selective pressure placed by drug treatment, subtypes display different mutation profiles (Quinones-Mateu et al., 2002; Vergne et al., 2000). Further, in response to drug treatment, they show evolution of clade-specific drug resistance mutations (Kinomoto et al., 2005; Spira et al., 2003).

It is now clear that equitable clinical management of patients presenting infections from different subtypes and/or CRFs may be of little value and therapy has to move towards genotype specific treatment regimes (Butler et al., 2007; Korber and Gnanakaran, 2009).

## 2.3 Genotyping HIV

To achieve effective clinical management of patients infected with HIV, it is imperative to be able to identify the infecting genotype successfully. The gold standard in genotyping has been phylogenetic analysis (Arens, 1999). However, these methods rely heavily on the choice of model and sequences used to construct the tree. Lack of experience in making these choices and in the interpretation of the resultant tree can therefore, easily lead to misguided classification. Over the past six years, some effort has been made to provide solutions to this problem that rely little on user expertise. The NCBI Viral genotyping tool (Rozanov et al., 2004) uses a reference database. A query sequence is then divided into overlapping windows and "BLAST"-ed against this database to produce a set of similarity scores. Visualized as a series of colored overlapping regions, the ancestry of the query may be assigned by the most similar reference sequence in that region. While a non-recombinant sequence is expected to present a monochromal similarity plot, recombinants will likely show blocks of colors separated approximately by the recombination breakpoints in the query. The Recombination Identification Program (RIP) (Gifford et al., 2006) uses a similar concept using a reference set to build a matrix of Hamming distances for overlapping windows of the query sequence. Once again, the reference sequence closest to the query in each window is designated as parent. Earlier, the STAR method (Gale et al., 2004) used a similar technique for sequences in the *pol* region.

The REGA genotyping tool (de Oliveira et al., 2005) genotypes a test sequence by

first using a fixed reference set to build a phylogenetic tree. It then uses measures of proximity of the query to a reference on the tree to genotype it based on a fixed binary decision tree. Wu et al. (2007) proposed a method of representing sequences as complete composition vectors (CCV) that capture the nucleotide makeup of a sequence based on observed and expected frequencies of oligomers. NCBI and RIP tools require further phylogenetic confirmation (Rozanov et al., 2004; Gifford et al., 2006) and "human" decisions to analyze the plots produced and genotype test sequences. CCV struggles with accurate genotyping of CRFs (Wu et al., 2007). Several independent studies have shown much disagreement among these methods (Gifford et al., 2006) and particular difficulties in identifying non-B subtypes and recombinant forms (Holguin et al., 2008b).

## 2.4 Machine Learning Algorithms

While there is high genetic diversity in HIV sequences, the above methods obviously suffer from not getting enough or the best information from this diversity to make accurate classifications. Our goal is to find a platform that can use simple summary data such as distances between sequences to discern latent patterns specific to sub-groups of the sequences, i.e. genotypes and use these to classify newer, more complex query sequences. Machine learning algorithms fit the paradigm of classification of complex data very well.

Specifically, we focus on supervised learning algorithms. The general setup of such algorithms involves the representation of the data we wish to classify, in the form of *features*. Using an example from above, the distance matrix of a query sequence against a reference set of sequences may be treated as a feature set. To achieve efficient classification, we require a **Training Set**. This is a dataset containing features for a set of queries for which the true classification is known. We provide the algorithm with this set and the list of true classifications. The algorithm then "learns" from this set, mapping

specific patterns to specific classification groups. It is then able to use features from a query sequence of unknown classification to look for known patterns from its "learning" step. Upon finding one such pattern, it successfully classifies the query.

Many algorithms have been proposed to optimize the learning and classification steps. Some popular ones do so by means of classification trees. A classification tree is a binary tree that poses a question at each internal node. The response to this question is binary and results in a query being sent down the left or right subtree. Ultimately, each query lands in one of the leaves of the tree, each of which represent a class. Algorithms vary on how they choose the splitting rules, i.e. questions in the internal nodes. Advanced algorithms use many trees instead of just one and differ in the way they summarize the classification from all the trees they grow.

### 2.4.1 Classification and Regression Trees (CART)

This is one of the earliest supervised learning algorithms (Breiman et al., 1984). As described above, it builds one binary tree from the training set. Split rules are chosen to best divide the data such as "Is value of column $j$ in the feature set $> x$ ?" where $x$ is some split value. Based on a binary response to this rule, each entity in the training/test set is assigned to the left subtree or right subtree. Recursively applied, this algorithm results in every entity traversing the tree and ending up uniquely in one of the many leaf nodes of the tree. Leaf nodes assign final classification/predicted regression values to member entities.

### 2.4.2 Random Forests

Many advance methods have been proposed since the introduction of CART. One such that applies a popular method called bagging to the CART method is that of Random Forests (RF)(Breiman, 2001). Bagging or Bootstrap aggregating results when $m$ bootstrap samples are obtained for a given training set. For each of these samples

a classification/regression tree is grown. Final classification is the mode of the $m$ trees in the case of classification and their average predicted value in case of regression. This method can thus be looked at as a special case of model averaging, also known as ensemble techniques. However, this method has been found to overfit some datasets, with noisy data accentuating the problem.

### 2.4.3 Bayesian Additive Regression Trees (BART)

BART is a technique that combines ensemble techniques with the idea of boosting. Boosting algorithms work on the premise that a set of *weak* learners can produce a single strong learner. BART is based on a *sum-of-trees* model (Chipman et al., 2008) which obtains the predicted response or classification as a sum over $m$ regression trees fit to a user provided feature set. Chipman et al. (2008) and Zhang and Hardle (2008) provide details of modifications that may be made to use the method for a classification paradigm. Suitable priors placed on the depth of trees insure that they remain weak, thus preventing over-fitting of the data.

### 2.4.4 Naïve Bayes algorithm

Methods exist that do not use the tree structure such as the popular Naïve Bayes algorithm (Domingos and Pazzani, 1997). It is based on a probability model that assumes independence of entities in the input feature set $\boldsymbol{X}$. The likelihood of the query sequence $j$ having response $Y_j$ is defined as,

$$p(Y_j|X_j) \propto p(Y_j) \prod_{k=1}^{n} p(X_{jk}|Y)$$

where $n$ corresponds to the number of columns in the feature set $X_j$. Priors and feature probability distributions are obtained from the training set. Classification models apply a decision rule such as the *maximum a posteriori* decision rule that picks the most probable hypothesis.

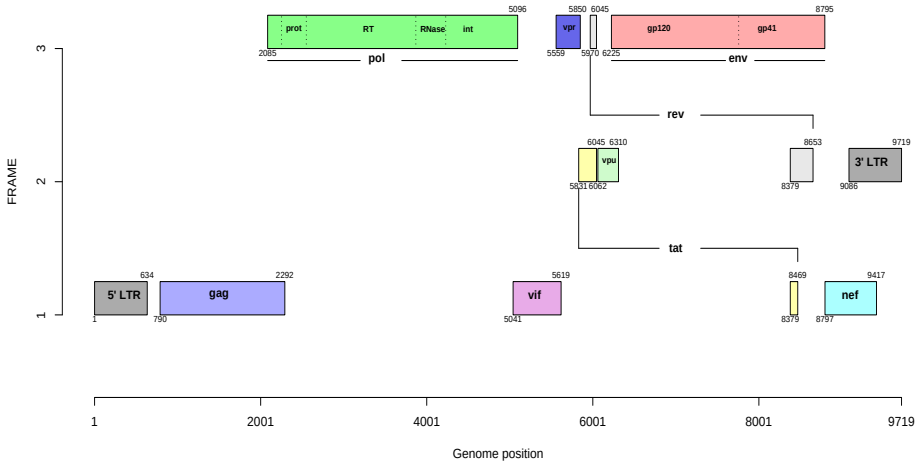



Figure 2.1 HIV genome organization

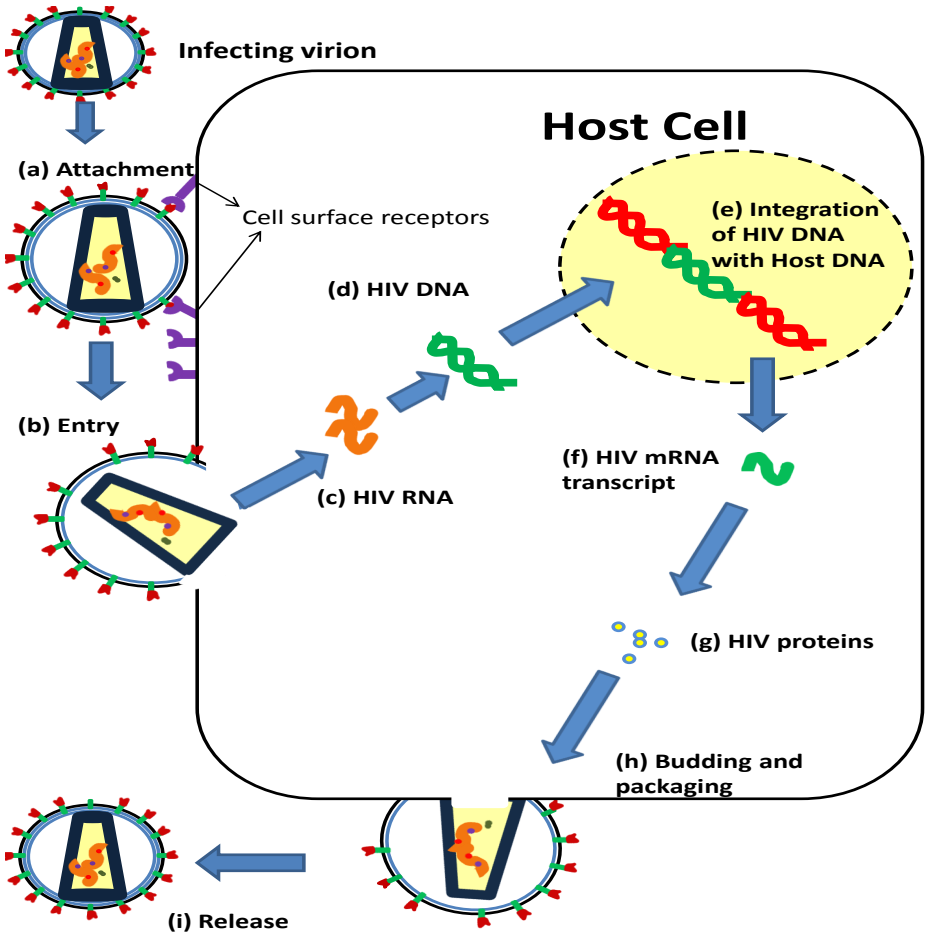



Figure 2.2 Major events in the replication cycle of HIV. (a) Free virus attaches to cell surface CD4 receptor and one of the co-receptors (CXCR4 or CCR5). (b) Fusion occurs upon successful attachment. (c) HIV RNA along with proteins necessary for replication, deposited in host cytoplasm. (d) Reverse transcription leads to generation of double stranded HIV DNA. (e) Integration of HIV DNA with host DNA. The virus is referred to as *provirus* at this stage and may lie dormant for many years before replicating. (f) Host transcription machinery produces HIV mRNA that is transported back to the cytoplasm. (g) Host translation machinery is used to produce HIV specific protein coded by HIV mRNA. (h) Viral proteins are packaged to form a new virion that buds off of the cell membrane of the host cell. (i) New virion is released from host cell into intercellular space.

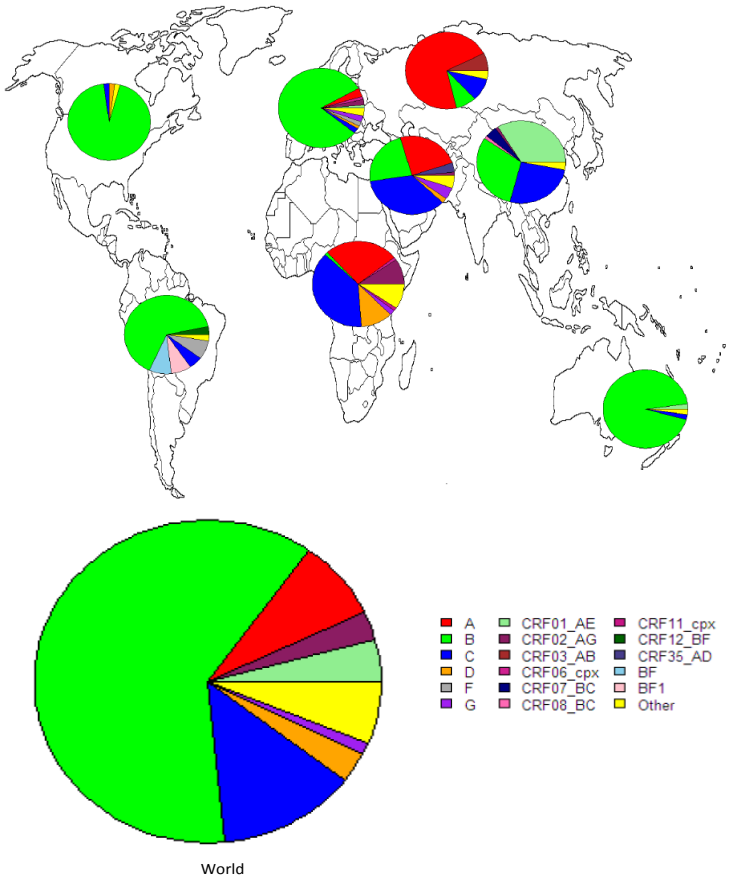



Figure 2.3 Global spread of HIV-1

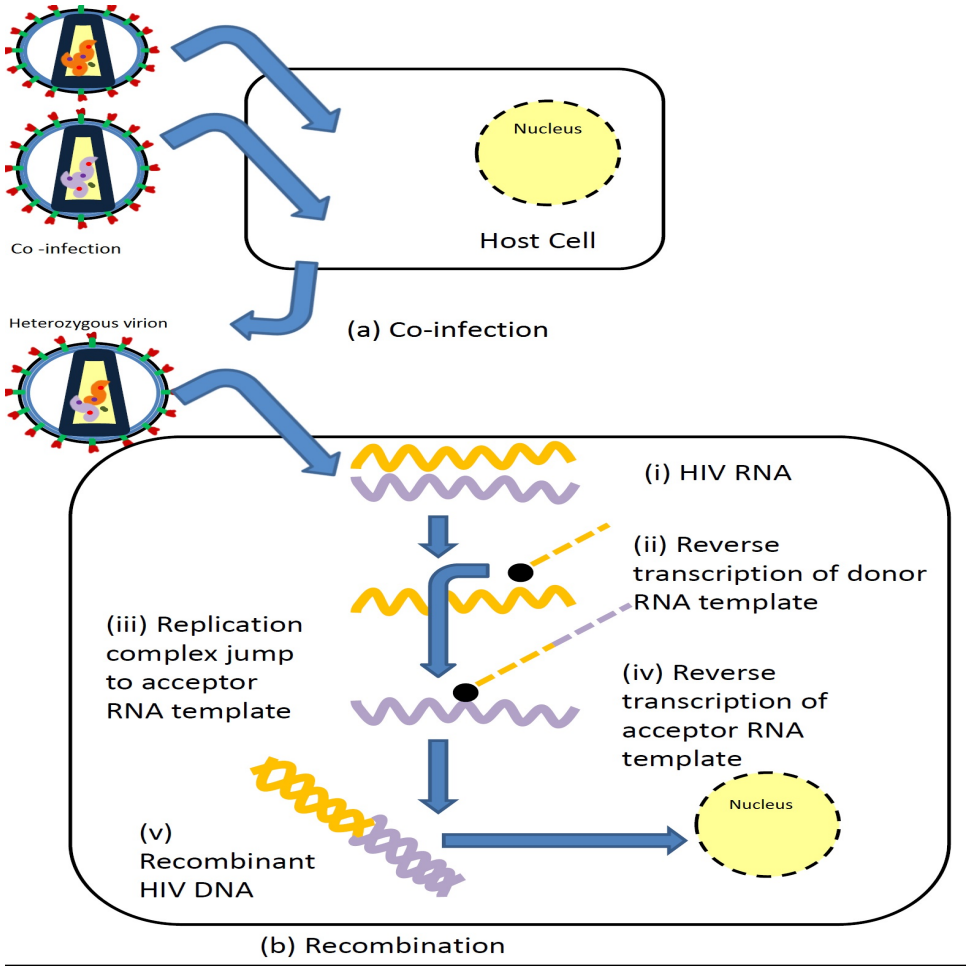



Figure 2.4 (a) Simultaneous infection of a single host cell with virions containing genetically distinct genomic material may lead to co–packaging of different genotypes within progeny virions. (b) Infection of a host cell with a virion co-packaging RNA genomes from distinct genotypes may lead to recombination. (i) HIV RNA is deposited in host cell following attachment and fusion of virion with host cell.(ii) Reverse transcription occurs using one of the two genomes as template. (iii) Transcription machinery "jumps" to the other genetically-distinct genome. (iv) Reverse transcription continues using the acceptor RNA as template. (v) Resultant double stranded HIV DNA is a mosaic of the two templates. Further replication takes place as described in Fig. 2.2 with the recombinant DNA.

# CHAPTER 3. RAPID GENOTYPING OF HIV-1 USING SUPERVISED LEARNING ALGORITHMS

A paper to be submitted to *Bioinformatics*

Misha L. Rajaram, Yves Sucaet, Karin S. Dorman

## Abstract

**Motivation:** The emergence of complex recombinant forms of HIV-1 and their documented fitness advantages over existing non-recombinant subtypes make it important to be able to efficiently genotype them. Existing methods do very well with non-recombinant types but face challenges with recombinant forms. Vast amounts of data are, however, available and may be used to advantage in order to achieve efficient genotyping.

**Results:** We present the application of supervised learning algorithms to the problem of genotyping. Sequences are represented either by a vector of oligomer frequencies or by measuring closeness to a reference sequence in an alignment using various measures. We demonstrate the method on a dataset of 500 sequences achieving respectively accurate classification. We also present a novel method, bootSMOTE, to generate synthetic sequences that may be used to supplement real data in cases of insufficient availability.

**Availability:** The software is available for download as well as through a web interface at `http://www.biomath.org/HIVgenotyping`.

## 3.1 Introduction

The Human Immunodeficiency Virus Type 1 (HIV-1) is currently believed to responsible for over 90% of global HIV infections (Kuiken et al., 2010). It is represented by 11 primary non-recombinant subtypes– A1, A2, B, C, D, F1, F2, G, H, J and K. Additionally there are 43 established recombinant subtypes known as Circulating Recombinant Forms (CRFs) and countless Unique Recombinant Forms (URFs) (McCutchan, 2006). Recombination occurs in retroviruses like HIV when the replication machinery switches templates from one of the packaged genomes to another. The resultant *recombinant* genome is a mosaic of the two originally packaged *parental* genomes. When the two packaged genomes belong to different genotypes this process results in the formation of inter-subtype recombinants (Negroni and Buc, 2001). Points on the recombinant genome across which the genome composition changes are known as recombination breakpoints. CRFs are inter-subtype recombinants found in three or more unrelated individuals showing the same recombination profile, i.e. breakpoints at similar locations and the same order of parental subtypes across breakpoints. Factors such as recombination, a high mutation rate as well as the complete absence of exonuclease proof reading contribute heavily to the genetic diversity of the virus (Spira et al., 2003).

CRFs, having triggered many local epidemics (Frange et al., 2008), now account for about 18% of the global epidemic (Buonaguro et al., 2007) and may gradually outnumber the pure subtypes in the global infecting pool (Tovanabutra et al., 2004; Toni et al., 2005; Njai et al., 2006; Wang et al., 2007). Subtypes and CRFs differ in virulence (Baeten et al., 2006), drug resistance (Spira et al., 2003; Madani et al., 2004) and sensitivity to detection assays (Swanson et al., 2005; Colson et al., 2007). Subtypes are also documented to differ in co-receptor preference with most subtypes favoring the CCR5 co-receptor in early disease and switching to CXCR4 after disease progression (Cornelissen et al., 1995). Subtypes A and C, favor the use of CCR5 throughout the infection time course (Peeters

et al., 1999), while subtype D utilizes both receptors (Kaleebu et al., 2002). Overall, knowledge of the infecting genotype is critical for efficient clinical management of infected individuals (Geretti, 2006).

The current gold standard method for genotyping is phylogenetic analysis (Arens, 1999) since it performs best especially in the case of complex recombinants (Holguin et al., 2008b). However, such analyses are time consuming and demand expertise in the meticulous choice and implementation of methods as well as interpretation of results. Over the past decade, some rapid and semi-automated genotyping procedures have been proposed (Rozanov et al., 2004; Gale et al., 2004; de Oliveira et al., 2005; Gifford et al., 2006; Wu et al., 2007). The NCBI Viral Genotyping Tool (Rozanov et al., 2004) and the Recombination Identification Program (RIP) (Gifford et al., 2006) use BLAST-based similarity scores of a test sequences against a set of reference sequences for an overlapping set of windows to determine the most similar reference "parental" type in each window. Similar to this approach, STAR (Gale et al., 2004) finds the closest sequence from a reference set of *pol* sequences. The REGA genotyping tool (de Oliveira et al., 2005) genotypes a test sequence using binary decision trees based on fixed reference sets. Wu et al. (2007) proposed a method for representing sequences as complete composition vectors (CCV) that summarize sequences based on observed and expected frequencies of oligomers. The NCBI and RIP tools require further phylogenetic confirmation (Rozanov et al., 2004; Gifford et al., 2006) and "human" decisions to genotype test sequences based on plots produced. CCV struggles with accurate genotyping of CRFs (Wu et al., 2007). Several independent studies have shown many discrepancies among these genotyping methods (Gifford et al., 2006) and difficulties in particular, identifying non-B subtypes and recombinant forms (Holguin et al., 2008b).

Rajaram and Dorman (2009) showed that of four popular supervised learning methods– Classification and Regression Trees (CART), Random Forests, Naïve Bayes and Bayesian Additive Regression Trees (BART)– BART performed best, especially

for complex HIV data that included unique recombinants. We present an ensemble HIV-1 genotyper comprised of several binary BART classifiers, one for each HIV-1 genotype. Together, these determine the genotype of the input sequence. We present results from using uncorrected distances or oligomer frequencies with both performing comparably on our test dataset. Further, we address some challenges supervised learning algorithms face when applied to HIV genotyping due to the skewness in available HIV data and present some solutions to deal with the same. A genotyping tool based on classification by BART is available for download as well as through a web interface at `http://www.biomath.org/HIVgenotyper`.

## 3.2 Methods

### 3.2.1 Classification with Bayesian Additive Regression Trees (BART)

Classification by supervised learning algorithms is achieved by the use of explanatory feature sets $X = (X_1, \ldots, X_l)$ to predict a categorical response $Y$. Feature sets are usually numerical summaries of a test entity that may help explain the response $Y$. These methods require a *training* set that provides feature representations for entities similar to the ones we wish to *test* or classify. Additionally, it is necessary to provide the true assignment of the $Y$ response variable for the training set. The learning algorithm then uses the training data to formulate "rules" that may be followed to obtain the classification for a test entity. Needless to say, the quality of the training data heavily impacts the success of a classifier. Additionally, the nature of the data may make some learning algorithms better for certain types of data than others.

BART achieves classification using a "sum-of-trees" model (Chipman et al., 2008) where the function that predicts $Y$ from $X$ is a sum of $m$ regression trees.

$$Y = \sum_{j=1}^{m} g(X; T_j, \mu_j) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

where each $g(X; T_j, \mu_j)$ is a classification/regression tree with structure $T_j$ and terminal node parameters $\mu_j$. A classification/regression tree has a binary structure. At each internal node, a "rule" splits the data arriving at the node, sending it down the left or right subtree depending on its value. Rules are formulated so they achieve the best split of the data. For example, a node may have a "variable $z > 0.9$" as a rule. Suppose that there are $n$ sequences/data points in the test set of which, $k$ satisfy this rule. Hence, the left subtree of the node will receive the $k$ data points that have a value greater than 0.9 for variable $z$ and the remaining $n - k$ data points will be passed down the right subtree. Trees are grown by recursively applying the above method. Applying the rules, each data point falls in exactly one leaf of each tree. The leaves of each tree contain the mean regression value for its constitueent data points. In BART, the final prediction is the sum of the regression values predicted by $m$ trees. BART guards against overfitting of data by placing a regularization prior on the height of each tree. This prior forces each tree to be a weak-learner thus preventing it from growing too tall and dominating the result. For classification of a binary response the BART model uses the probit link function (Chipman et al., 2008; Zhang and Hardle, 2008). The predicted outcome is in the form of a continuous variable that is thresholded at 0.5 to make a binary classification.

### 3.2.2 Feature Representation

Feature representations used in the current study can be broadly classified into alignment based features and non-alignment based features. Alignment based features are obtained by characterizing the relatedness of a sequence to a reference representative of a genotype of interest from pairwise sequence alignments involving the two. On the other hand, non-alignment based features simply decompose every sequence into a set of numerical coordinates and use the resultant vectors to compare with the vectors obtained members of the training set to classify. It must be noted that these are only a few

types of representations and may be changed for other representations that may yield better prediction for specific sets of sequences.

**Alignment based features:** The alignment-based features used in this study summarize a sequence in terms of its "similarity" with parental genotypes. In order to have a standard set of parental genotype sequences to which to compare and generate these features, we curated a **Reference Set**. The 2009 complete genome reference sequences was downloaded from the HIV database (HIV-Database, 2010). We reduced this reference set to contain one representative sequence per subtype/CRF by using consensus sequences when three or more representative sequences were present. When only two representative sequences were available, one was chosen randomly. Because recombinants are expected to differ in relatedness to a single reference genotype across their genome, a global full length measure does not quantify the relatedness metric accurately. To address this concern we used a sliding window based method to obtain a vector of pairwise relatedness measures for a query sequence with each reference genotype.

The simplest measure of relatedness is uncorrected distance. The $UD$ feature set refers to the collection of vectors specifying the relatedness of query sequences to a reference genotype. Specifically, $UD_j$ refers to a matrix storing the uncorrected distance of all query sequences against a specific genotype $j$, with each query sequence represented by a row and columns representing consecutive windows.

**Non-alignment based features:** Alignment based features, while very effective (Rajaram and Dorman, 2009) come with the overhead of having to build alignments. This can be very time consuming for a large batch of sequences. Although, in the case of training sequences, this may only be a one-time investment. We borrowed from the idea of Wang et al. (2007) and summarized each sequence as a vector of frequency of occurrence of certain oligomers. The reference set was used with the CCV program described

in Wang et al. (2007) to obtain a list of the top 5000 most informative oligomers ranging from size 7 to 21 bp. Counts of occurrence of each of these 5000 oligomers constituted the $Kmer$ feature set.

### 3.2.3 Classification with Binary Classifiers

We achieved genotyping using a set of one-versus-all binary classifiers. For each genotype we wished to test, we built a binary classifier. The classifier for genotype $j$ trained from feature representations of training sequences that had response $Y_i = 1$ if their true genotype assignment contained at least one fragment of genotype $j$ and 0 otherwise. Trees trained thus were then used to classify test sequences for which $Y$ values are unknown to the classifier. Final genotyping results were obtained by listing all the genotypes for which a particular test sequence got a positive classification. Fig. 3.1 describes the binary classification process.

### 3.2.4 Cross-validation

In order to test the efficacy of our classifiers and to make sure that they are not inordinately affected by choice of training and test data, we performed a 5-fold cross validation test. At each fold, 4/5ths of the data was used to train a classifier and the held-out 1/5th was used to test it. Denoting true positives (negatives) $TP$ ($TN$) and false positives (negatives) $FP$ ($FN$), $P = TP+FN$, $N = FP+TN$, $P' = TP+FP$, and $N' = TN+FN$, we report accuracy $(TP+TN)/(P+N)$, specificity $TN/N$, sensitivity $TP/P$, precision $TP/P'$, Matthew's correlation coefficient ($MCC = \frac{TP \times TN - FP \times FN}{\sqrt{P \times N \times P' \times N'}}$) and F-measure ($F = \frac{2 \cdot \text{precision} \cdot \text{sensitivity}}{(\text{precision}+\text{sensitivity})}$). Precision, MCC and F-measure are recommended assessments for imbalanced datasets (Chawla, 2006).

## 3.3 Data

### 3.3.1 Full length Dataset

To create a relevant dataset for an evolving virus now sequenced for several decades, we chose the most recent 500 sequences. Genotypes with less than 5 full length genome representatives had to be eliminated from this dataset in order to be able to use it for a 5-fold cross validation study, leaving us with 18 genotypes. To insure fair representation of all 18 genotypes, we randomly substituted highly-represented genotypes with older, under-represented genotypes, until there were at least 5 full-length genomes for all 18 genotypes. The final dataset (shown in Table 3.6) included 337 pure subtype nonrecombinants, 34 URFs of pure subtypes, 112 CRF sequences and 17 URFs with at least one CRF parent. Datasets and accession numbers can be found at
`http://www.biomath.org/HIVgenotyping`.

### 3.3.2 Gag Dataset

This dataset consists of all publicly available *gag* sequences representative of HIV-1 group M (HIV-Database, 2010). Of these we randomly chose 100 sequences to form a *Test Gag set* while the remaining $3,937$ were used to form the *Training Gag Set*.

### 3.3.3 Full length Supplemented Dataset

This dataset was curated to be used to train trees on all genotypes for the full length HIV genome. Starting with all $1,776$ publicly available full length HIV-1 sequences as of September 2008, we supplemented all under-represented CRFs and subtypes with 10 synthetic non-recombinant and 10 synthetic recombinant sequences (method described in later sections). The resultant dataset had $2,676$ sequences.

## 3.4 Challenges

### 3.4.1 Dealing with Imbalanced Data

It is obvious from a glance at Table 3.6 that the distribution of available data by genotype is heavily skewed. In the case of classification, this creates a problem, especially for classifiers of under-represented genotypes. Consider a case where all known full-length genomes are used to train a classifier for genotype CRF42_BF. The training set will then have 17 positive cases versus $1,759$ negative cases. Metrics of accuracy are misleading in such cases since predicting all cases as negative still gives the classifier a 99% accuracy. Provost (2000) and Chawla (2006) point to adjusting thresholds as a means of dealing with such situations. Instead of fixing the threshold at 0.5 for a classification problem, we optimize it, so as to minimize errors and achieve correct classification of the positive cases as well.

### 3.4.2 Supplementing Data with Synthetic Recombinants: bootSMOTE

For emerging recombinant forms, however, we are further hampered by having very little information. For many CRF types we have two or three sequences that also tend to have very little diversity. Training on one or two and testing on the remaining therefore works well currently but is not expected to deal well with future diversity. Unique recombinant forms, as also newer CRFs involving these heavily underrepresented genotypes, may end up misclassified due to lack of training data. In order to provide a little more balance to the training data we propose to supplement with synthetic data that we generate using a novel technique, bootSMOTE. This method uses the general idea of a bootstrapping and applies it as a Synthetic Minority Oversampling TEchnique (SMOTE) proposed by Chawla et al. (2002) to deal with imbalanced datasets. Synthetic variants were created using $M \geq 2$ "real" non-URF sequences of a genotype in a multiple sequence alignment of length $N$. For each position $l \in 1, 2 \ldots N$ of the new variant, a

"real" sequence $j \in 1, 2 \ldots M$ was chosen at random and the nucleotide in its position $l$ was used. The resultant set of variants remained close enough that they formed a monophyletic group with their respective real sequences. At the same time, they had more variation, albeit synthetic, that may better represent the variation in hitherto unseen members of the genotype.

To produce a synthetic recombinant, the number of breakpoints was considered to be a Poisson random variable with a user defined mean value (e.g. a mean of 0.3 breakpoints produces 80% nonrecombinant sequences and tends to limit the number of breakpoints to a maximum of two). Conditional on the number of breakpoints, their location was allowed to be uniform over the length of the entire genome ensuring that each fragment was at least 200 bp long. For each fragment, a genotype was chosen at random constraining only for chosen genotypes to be different across a breakpoint. Individual fragments were then constructed using the bootSMOTE method described above using non-URF training sequences of the chosen genotype. Figure 3.2 provides a schematic representation of the bootSMOTE technique.

### 3.4.3 Using Partial Genome Trees

Another type of paucity of data arises when faced with partial genome sequences. It is important to examine if and how missing data, due to lack of sequence information of a test sequence affects performance by a classifier that has been trained on longer sequences. We trained trees using the *Gag Training set* and feature sets $UD$ and $Kmer$. We also trained trees for the teo feature sets using the *Full length supplemented set*. The *Gag Test set* was used to test all these classifiers. Further, we truncated the 100 sequences in the *Gag Test Set* by 250bp and 500bp at the 3' end and used the *gag* trained trees to classify these.

### 3.4.4 Software

The source code for the genotyper contains a Perl wrapper script that calls the classification routine from a C++ library. The C++ library is a modified version of the BART R package (Chipman et al., 2008). The code may be used to train and store classifiers using a training set of sequences. Alternately, stored classifiers may be provided to the code to be used for classification of test sequences. A web interface for the HIV-1 genotyping tool can be found at `http://www.biomath.org/HIVgenotyper`. This interface has a PHP front end and the Perl–C++ code at the back end. The web form accepts test sequences from users as raw sequences as well as feature sets. Also, it provides the user with pre-made trees that may be used for classification. Results are displayed in simple text format.

**Pre-made classifiers:** Currently, the genotyper provides two sets of pre-made trees, both trained on the *Full length supplemented dataset*. For the first set, the *UD* feature set was computed using the *Reference Set* and used to build the training trees. The second set was built using feature set $Kmer$ corresponding to the training set. Having pre-made trees makes classification of new test sequences virtually instantaneous since it does away with the time overhead of training.

## 3.5 Results

For alignmemt based features, we used window sizes of 300 bp placed every 100 bp. This choice is arbitrary and any choice is suitable as long as there is considerable overlap between successive windows (data not shown). Table 3.2 provides details of performance of classification using various combinations of methods and feature sets. Performance measures reported are mean values across the 5-fold cross-validation runs with standard errors in parentheses. Holguin et al. (2008a) found the NCBI viral genotyping tool to

be better than STAR and REGA in classifying CRF06_cpx sequences. In the current analysis we found that all three methods, NCBI, REGA and CCV were able to classify pure subtype sequences almost faultlessly, as was our genotyping tool, regardless of the feature set that was used. REGA and NCBI had difficulties specially when confronted with CRFs or URFs that are not part of the reference set. NCBI allows for a user submitted reference set and performance improved for CRFs with a more relevant reference set while URFs were still hard for the NCBI tool to classify. The CCV method, on the other hand, had some difficulty with classifying CRFs even when they were in its reference set and had very limited success with complex URFs. For both the NCBI and CCV output, the closest $n$ predicted parental types were taken as the classification provided by the method where $n$ is the number of real parental genotypes of the test sequence. BART handled CRFs present in the reference dataset with ease. Both the $Kmer$ and $UD$ datasets perform comparably. Additionally, our genotyper was able to correctly classify most URFs except in cases where a fragment was particularly small.

### 3.5.1 Using Synthetic Recombinants

Training and test datasets were created to simulate the current state of data availability on certain genotypes. Some have only two representative sequences and it is possible, with the fast evolving nature of the virus, that newer sequences involving these genotypes evolve. In cases such as these, the training data available is insufficient for efficient classification of emergent complex forms. To test such scenarios we used, as an example, the subtype D classifier trained on feature set *UD*. In the *Full length dataset*, subtype D is represented by two non-recombinant sequences and four URFs containing fragments of subtype D. While the non-recombinants were identified correctly, the classification of the URFs suffers from lack of training information. Fig. 3.3 demonstrates the use of bootSMOTE synthetic recombinants to enhance classification. We generated 50 synthetic non-recombinant sequences using the two non-recombinant D sequences

already present in our dataset. However, supplementing with these made no difference to identification of the D fragment in the four URFs (black line in Fig. 3.3).

Not included in our current dataset are 54 non-recombinant D sequences that were eliminated in favor of more recent representatives. We chose three of these randomly to see if the identification of URFs improved. Supplementing with just the non-recombinant sequences results in two URFs being identified (blue line is Fig. 3.3). However, supplementing with 10 synthetic sequences generated from three non-recombinants leads to identification of one URF (red line in Fig. 3.3). This performance is matched with only four synthetic nonrecombinants generated from four real sequences.

To test if supplementing with synthetic recombinants improves performance we generated 50 synthetic recombinants with at least one fragment belonging to the D genotype generated using only the 2 nonrecombinant sequences that were part of the *Full length dataset*. Supplementing with 30 synthetic recombinants results in identification of the two URFs previously identified.

When real data is available in the form of only two sequences, the similarity of the available sequences is bound to affect any synthetic sequences generated from them and consequently classification achieved by training on these synthetic sequences. To study the effect of the level of similarity of real sequences we started with four pairs of non-recombinant B sequences with varying pairwise uncorrected distances ranging from 0.0 to 0.1451. Fifty synthetic recombinants were generated from each pair. Each synthetic recombinant had at least one fragment belonging to subtype B, generated from the original pair of B sequences. Classification of the synthetic recombinants was attempted by using just the real B sequences as training data and later supplementing the training data in steps of 10 synthetic sequences. Table 3.6 presents results from this analyses. For closely related sequences, there was a high success rate of true positive identification with this rate reducing to almost 0 with increasing distance between the pair of parents, when the $UD$ feature set was used. In the case of using the $Kmer$ feature set, even

closely related parental pairs only result in a true positive rate of $\sim$ 50% although we still see the trend of smaller true positive rates with increasing distance between parents. Interestingly, supplementing with 10 synthetic recombinants improves this rate considerably in all cases. To complete the picture, we used recombinants created using pair one as a test set while using pair four as the training set. Supplementing the training set with 10 recombinants made from pair four improved the true positive identification from 80 to 96% in the case of using the $UD$ feature set and from 48 to 100% in the case of the $Kmer$ feature set.

### 3.5.2 Partial Genome Trees

It is common for researchers to sequence only a part of the genome. In cases like this, using classification trees trained on longer data may result in poor classification. To test this premise, we classified sequences in the *Gag Test Set* using the pre-made full length training trees described in section 3.4.4 as well as *Gag Train Set* with feature sets $UD$ and $Kmer$. Table 3.6 shows results of classification success with both these training scenarios using both the $UD$ and $Kmer$ feature sets. While using trees trained on *gag* sequences, classification using the $UD$ feature set is comparable to earlier results in Table 3.2. However, using trees trained on full length data, affects the outcome considerably. Interestingly, with the $Kmer$ feature set, even trees trained using *gag* sequences have somewhat lowered classification power. Table 3.6 also shows classification performances of $UD$ trees trained on the *gag* sequences while classifying truncated *Gag Test* data. Classification performance rapidly decays as the proportion of missing data increases.

## 3.6 Discussion

In this study, we presented a novel application of supervised learning algorithms to the very pertinent problem of rapid HIV-1 genotyping. In order to deal with recombinants, we implemented a dedicated binary classifier for each genotype of interest so that a test sequence has the opportunity to be classified simultaneously within several genotypes. This approach leads to near accurate classification of not only known CRFs but URFs as well.

The skewed sampling of HIV-1 sequences, with some genotypes represented more than others and some regions of the genome sequenced more than others. This leads to a variety of situations that provide little or no data, necessitating the exclusion of some genotypes from the ensemble genotyper. In order to include genotypes with insufficient data and hence build a more complete genotyper we proposed supplementation with synthetic sequences generated using the bootSMOTE method. As illustrated using the case of genotype D, the classifier accurately identified the two full length sequences, however, failed to classify the four URFs containing genotype D also in the *Full length dataset*. In an evolution of the virus, with more and more recombinants surfacing as viable pathogens, training on just non-recombinant sequences may not work well. We encounter this scenario in the example of the D classifier where even when supplemented with fifty synthetic sequences generated from the two non-recombinant D sequences, none of the URFs were identified. However, addition of just one more non-recombinant D sequence to the data used to generate synthetic variants, leads to success in identifying one URF. However, in most real cases, we may only have two or three sequences. In such cases, this example also demonstrated the power of using synthetic recombinant sequences for supplementation. These synthetic recombinants, even when originating from just two non-recombinant sequences lead to successful identification of two URFs. In two of the four URFs the D fragment still remains unidentified. It is possible that

the fragments are too small for detection with our current feature sets.

We explored the power of supplementation with synthetic recombinants more with the experiment involving pairs of B sequences. The number of synthetic sequences required for fully accurate identification of all recombinants in the test set increased with increasing pairwise distance between the original pair of real sequences. However, all recombinants were eventually identified successfully. Synthetic recombinants supplement data in many ways. In the case where the pair of B sequences were very close (pairwise uncorrected distance = 0.0), the B fragments of all synthetic recombinants will naturally be very similar to each other as well as the parental sequences. However, even with such striking similarity, the classifier was not able to correctly classify all of them when trained only on real data. This is possibly because recombinants lack the *complete signature* that the trained classifier would have identified in the real data. Including recombinants in the training set allows the classifier to calibrate to instances where only a fragment in the entire sequence belongs to the B genotype i.e. synthetic recombinants inform on structure. Cases with divergent real data need more help. In these cases, not only do the synthetic recombinants inform on structure, they now also make up for paucity in content. Diversity of the parental sequences has direct bearing on the diversity in the synthetic recombinant. Thus, when tested with one such synthetic recombinant, the classifier has to identify a fragment and classify as belonging to the B genotype. The dual information provided by the synthetic recombinants in such cases is reflected in the increasing number required to classify all the test recombinants correctly.

The use of specialized training sets is a strength of this method. As illustrated in the example using partial genome trees, missing sequence data can be highly detrimental to efficient classification. Using training sequences that are comparable in length and covering the same genomic region produces best results. Both feature sets perform well, albeit, a little worse than their classification performance on the *Full length dataset*. The short length of the region may yield much less information for the classifier to train on.

Attempts at using classifiers trained on full length sequences produced poor results. In order to use these classifiers, we disregarded all trees that contained split rules involving missing data. This greatly reduces the number of trees used in the final classification and potentially loses the granularity required to distinguish a recombinants from its non-recombinant parent. Similar results are obtained when the classifiers trained on the *Gag Training dataset* where tested using truncated *gag* sequences.

The adaptability of this genotyper to the use many different relatedness measures and ability to incorporate additional sources of information or very large sets of covariates is a big asset. Rajaram and Dorman (2009) demonstrated the use of two other alignment based features that performed nearly as well as the $UD$ feature set. It is possible to provide other feature sets such as distances to genotypes other than the target genotype (preliminary analysis shown in Rajaram and Dorman (2009)) or even bootscan results (Salminen et al., 1995). Further, multiple information such as distance and phylogenetic measures could be combined.

Phylogenetic based recombination detection tools (Gale et al., 2004; Husmeier and McGuire, 2003; Grassly and Holmes, 1997; Suchard et al., 2003; Minin et al., 2005), while successful, are slow particularly in the case of complex recombinants that may require inclusion and analysis of multiple putative parental genotypes. Though a more thorough window-by-window application of supervised learning algorithms could provide an estimate of recombination breakpoints, we envision this method as a rapid screening tool that may be followed up by phylogenetic detection analyses. In such a case, this tool can greatly decrease the list of putative parents of complex recombinants thereby shortening the time taken for downstream analyses. While training classifiers takes time, it is modest compared to phylogenetic analyses and is a one-time cost. The classification of test sequences using pre-made classifiers is virtually instantaneous. Analysis of a test sequence with BART-UD takes about eight seconds on a modern 64 bit server; REGA and NCBI were reported to take about 10 seconds per test sequence (de Oliveira et al.,

2005).

Users may use the source code to train and store classifiers most pertinent to the test data they wish to classify. For example, reference and training sets specific to a geographical region may perform better classification of more sequences from the same region. The web interface provides users with the opportunity of rapidly genotyping their test sequences using pre-made classifiers. It is proposed to add more pre-made classifiers to the interface pertaining to specific regions of the genome as well as different feature sets so as to make the interface more useful to a larger user group.

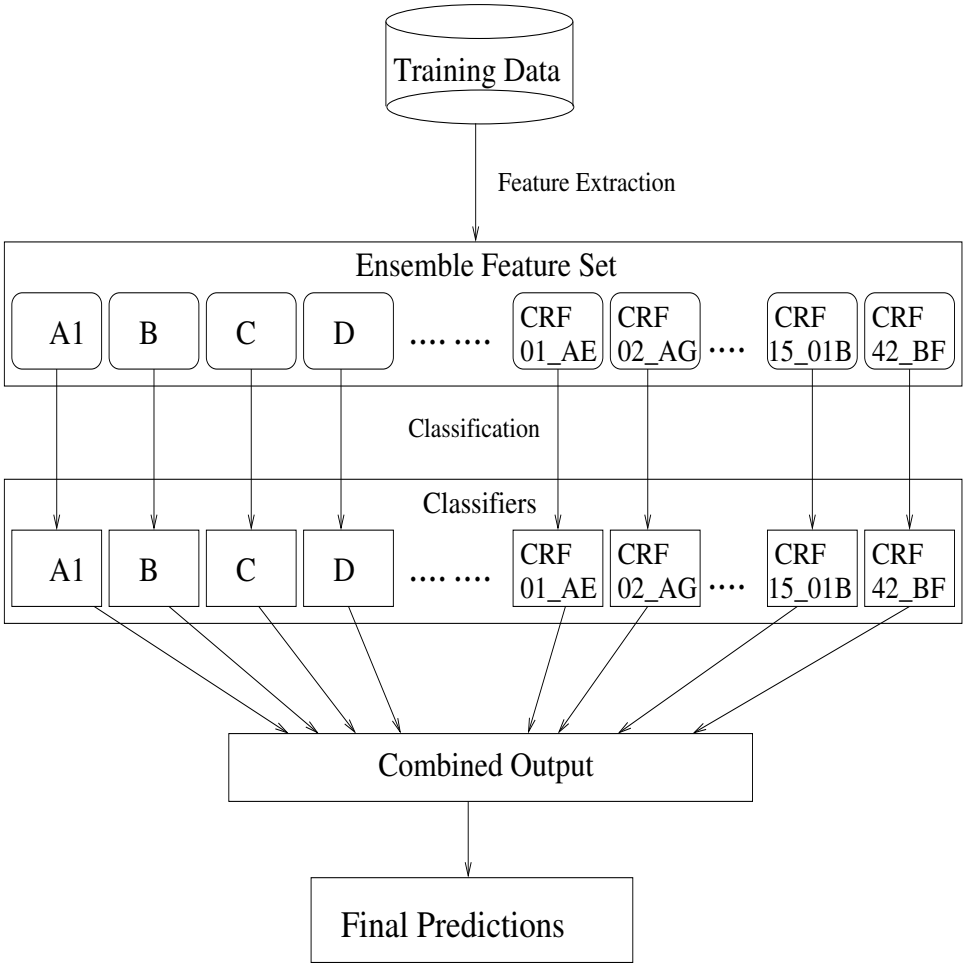
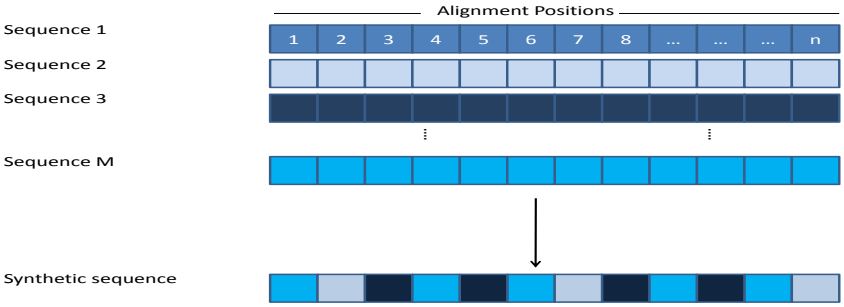
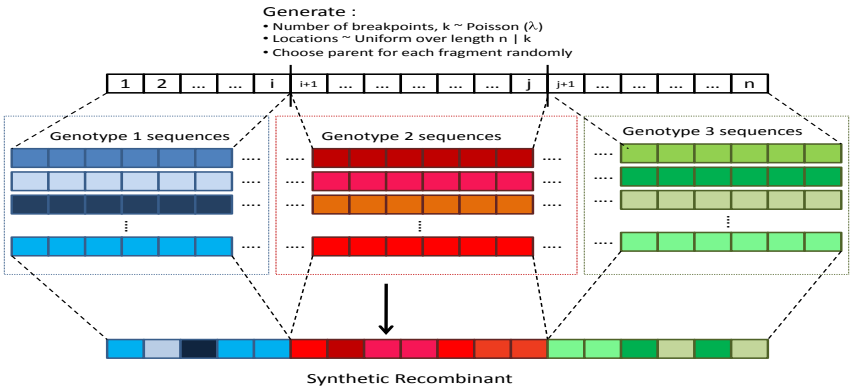



Figure 3.1 The ensemble genotyper setup with a binary classifier representing each genotype. Final classification is a list of all positive classifications from the ensemble.




(a) Creating synthetic non-recombinants




(b) Creating synthetic recombinants

Figure 3.2 (a) bootSMOTE for creating synthetic non-recombinants starting with an alignment of M real non-recombinant sequences. At each position $1 \leq i \leq n$ one sequence from the alignment is chosen randomly and used in synthetic sequence. (b) bootSMOTE procedure for creating synthetic recombinants. Here, $k = 2$ breakpoints are placed at positions $i$ and $j$. Parental genotypes are chosen for each fragment. Synthetic recombinant is generated by concatenating fragment sequences generated as shown in (a).

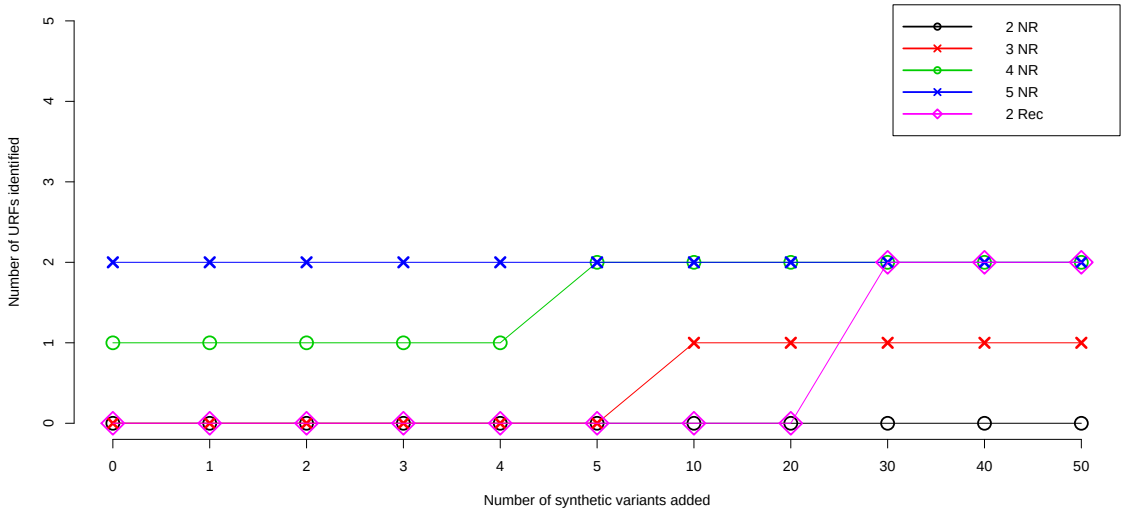



Figure 3.3 Classification of URFs by subtype D classifier upon supplementation. Plot showing number of URFs identified with addition of synthetic variants in steps (x-axis). NR (Rec) denotes use of nonrecombinant (recombinant) variants for supplementation. The number 2, 3, 4 or 5 denotes the number of real D sequences used to create synthetic variants.

| Genotype | U | S/C | Genotype | U | S/C |
|---|---|---|---|---|---|
| CRF 01_AE | 10 | 20 | CRF 02_AG | 6 | 27 |
| CRF 06_cpx | 2 | 7 | CRF 07_BC | 1 | 5 |
| CRF 08_BC | 0 | 5 | CRF 09_cpx | 1 | 5 |
| CRF 11_cpx | 0 | 6 | CRF 12_BF | 0 | 5 |
| CRF 13_cpx | 0 | 5 | CRF 14_BG | 0 | 5 |
| CRF 15_01B | 0 | 5 | CRF 42_BF | 0 | 17 |
| A1 | 13 | 22 | B | 33 | 74 |
| C | 21 | 235 | D | 4 | 2 |
| F1 | 12 | 1 | G | 3 | 3 |

Table 3.1 Distribution of genotypes in full length dataset, showing counts of URF (U), pure subtype/CRF (S/C) for each

| Method | Acc. | Spec. | Sens. | Prec. | MCC | F |
|---|---|---|---|---|---|---|
| **REGA** | 0.97 | 0.99 | 0.61 | 0.99 | 0.76 | 0.754 |
| **NCBI** | 0.80 | 0.78 | 0.98 | 0.25 | 0.43 | 0.398 |
| **CCV** | 0.924 | 0.96 | 0.387 | 0.387 | 0.346 | 0.387 |
| **BART-Kmer** | 0.990(0.002) | 0.997(0.002) | 0.919(0.021) | 0.960(0.030) | 0.934(0.017) | 0.939(0.015) |
| **BART-UD** | 0.991(0.002) | 0.995(0.001) | 0.929(0.019) | 0.922(0.017) | 0.923(0.017) | 0.926(0.015) |
| **CART-UD** | 0.988(0.001) | 0.994(0.0003) | 0.899(0.022) | 0.914(0.004) | 0.900(0.011) | 0.906(0.011) |
| **RF-UD** | 0.985(0.002) | 0.976(0.013) | 0.986(0.001) | 0.888(0.016) | 0.821(0.018) | 0.891(0.017) |
| **NB-UD** | 0.800(0.015) | 0.963(0.014) | 0.789(0.017) | 0.415(0.0191) | 0.232(0.014) | 0.543(0.014) |

Table 3.2 Performance of classifiers. Mean (and standard error) of performance measures across 5-fold cross-validations.

| | Pairwise Distance | # suppemented recombinants | TPR (UD) | TPR(Kmer) |
|---|---|---|---|---|
| Pair 1 | 0.00 | 0 | 0.96 | 0.52 |
| | | 10 | 1 | 1 |
| Pair 2 | 0.04306 | 0 | 0.84 | 0 |
| | | 10 | 0.96 | 0.92 |
| | | 20 | 1 | 1 |
| Pair 3 | 0.07003 | 0 | 0.04 | 0 |
| | | 10 | 0.76 | 0.56 |
| | | 20 | 1 | 1 |
| Pair 4 | 0.14512 | 0 | 0 | 0 |
| | | 10 | 0.6 | 0.8 |
| | | 20 | 0.94 | 0.98 |
| | | 30 | 1 | 1 |

Table 3.3 Effect of relatedness of parental sequences on synthetic recombinant supplementation. Table presents true positive rate (TPR) of identification of synthetic B recombinants with classiers trained with a pair of real B sequences of varying pairwise distances as well as results with classifiers trained on data supplmented with synthetic recombinants.

| Training Tree | Acc. | Spec. | Sens. | Prec. | MCC | F |
|---|---|---|---|---|---|---|
| Gag - UD | 0.967 | 0.983 | 0.832 | 0.852 | 0.824 | 0.842 |
| Gag - Kmer | 0.964 | 0.988 | 0.771 | 0.895 | 0.811 | 0.828 |
| Full - UD | 0.582 | 0.559 | 0.8 | 0.160 | 0.211 | 0.267 |
| Full - Kmer | 0.573 | 0.545 | 0.817 | 0.173 | 0.221 | 0.286 |
| Analyses of truncated Gag Test Sets using the Gag-UD training trees | | | | | | |
| Test Set | Acc. | Spec. | Sens. | Prec. | MCC | F |
| Gag - 250bp | 0.537 | 0.524 | 0.643 | 0.136 | 0.103 | 0.225 |
| Gag - 500bp | 0.491 | 0.489 | 0.510 | 0.131 | −0.0004 | 0.209 |

Table 3.4 Classification performance using classifiers trained on Gag sequences and full length sequences and tested using Gag test sequences.

# CHAPTER 4. HOT AND COLD: SPATIAL FLUCTUATION IN RECOMBINATION RATES

A paper published in the *Proceedings of the 7th Annual IEEE Bioinformatics and Bioengineering Conference,2007*

Misha L. Rajaram, Vladimir N. Minin, Marc A. Suchard and Karin S.Dorman

## Abstract

Coinfection of a single cell with two or more HIV strains may produce recombinant viruses upon template switching by the replication machinery. We applied a hierarchical multiple change point model to simultaneously infer inter-subtype recombination breakpoints and spatial variation in the recombination rate along the HIV-1 genome. We examined thousands of publicly available HIV-1 sequences representing the worldwide epidemic and focused on 544 unique recombinants with $1,701$ recombination breakpoints. Estimates of per site recombination rate revealed the presence of a novel hotspot in the *pol* gene, surrounded by a cluster of mutations associated with resistance to reverse transcriptase inhibitors. We also confirm the presence of a known hotspot in the *env* gene and a previously hypothesized hotspot in the *gag* gene.

## 4.1 Introduction

As if to achieve a primitive kind of sexual reproduction (Temin, 1991), retroviruses, including Human Immunodeficiency Virus (HIV), package two positive sense RNA parti-

cles and a strand-switching reverse transcriptase enzyme in each virion. Even genetically distinct strains can be copackaged (Hu and Temin, 1990), and with documented cases of superinfection of both hosts (Chohan et al., 2005; Fang et al., 2004; Hu et al., 2005) and cells (Levy et al., 2004), all ingredients necessary for recombination to mold the AIDS epidemic are in place (Peeters, 2000). Here we focus on the *in vivo* spatial distribution of strand transfer events along the genome to learn about the mechanism and selection of recombination in HIV type 1. The error-prone reverse transcriptase (RT) has produced pronounced variation in HIV (Wain-Hobson, 1993). Viruses around the world have been classified into subtypes (e.g. A), circulating recombinant forms (e.g. CRF01_AE), and subsubtypes (e.g. A1) (Robertson et al., 2000). All major subtypes have long co-circulated in Africa, but subtype mixtures are increasingly common in other parts of the world (e.g. (Bello et al., 2007; Cuevas et al., 2002)). With few barriers to limit recombination between genetically diverse strains (Baird et al., 2006b; Chin et al., 2007), the prevalence of recombination seems destined to increase as the epidemic progresses. The capacity to recombine diverse viral strains is almost certainly a benefit to retroviruses (Burke, 1997; Worobey and Holmes, 1999).

HIV recombinants have successfully out-competed other strains within hosts (Fang et al., 2004; Salminen et al., 1997), and inter-subtype recombinants have been alarmingly successful at causing (e.g. (Liitsola et al., 1998; Piyasirisilp et al., 2000)) or taking over (Tee et al., 2005) local epidemics.

With recombination thus driving HIV evolution, it is important to uncover the mechanism of strand transfer. Increasing experimental and observational evidence suggests that strand transfer events do not occur uniformly along the genome. There have been several hints of a hotspot in the 5' portion of the *pol* gene, both in experiment (Jetzt et al., 2000) and among *in vivo* recombinants sampled from patients (Magiorkinis et al., 2003; Thomson et al., 2004). Another well studied hotspot is the conserved C2 region of *env* (Moumen et al., 2001; Quinones-Mateu et al., 2002), although all conserved regions

in *env* are relatively hot (Baird et al., 2006a) and even variable regions can become hotspots with the right donor template (Baird et al., 2006b). Many inter-subtype recombinants observed around the world display a recombinant pattern where the variable loop V3, between C2 and C3 (Renjifo et al., 1999), or the complete gp120 portion of *env* (Takebe et al., 2003) is swapped with another subtype. Other regions implicated as possible hotspots are 5' *gag* (Dykes et al., 2004), the *gag-pol* boundary (Magiorkinis et al., 2003), the *pol-vif* boundary (Derebail et al., 2003), through *vif* into the 5' *env* (Magiorkinis et al., 2003), a GC-rich region near the *tat/rev* splice site (Douglas et al., 1996), and near *nef* (Magiorkinis et al., 2003). Indeed, very few genomic regions are consistently "cold," but few studies have examined the entire genome at once and experimental protocols and reagents vary widely.

There is no single determinant of retroviral recombination, but at least secondary structure, reverse transcriptase pausing, and sequence homology influence transfer rates (see recent review (Galetto and Negroni, 2005)). The recombination enhancing role of nucleocapsid (Negroni and Buc, 2000) suggests the importance of secondary structure, and the TAR stem loop is one structure known to facilitate strand transfer (Berkhout et al., 2001; Kim et al., 1997). Recombination in the *env* C2 region depends on homology between the donor and acceptor templates, a stable stem loop in the acceptor, and to a lesser degree, primary sequence (Galetto et al., 2004, 2006; Moumen et al., 2003). Homology at the dimerization signal (Chin et al., 2007; Moore et al., 2007) and throughout the genome enhances inter-molecular strand transfer (Andersen et al., 2003), most importantly at the strand transfer site (Baird et al., 2006b; Gao et al., 2007). It was recognized early that strand transfers tend to occur at pause sites of reverse transcription (Wu et al., 1995; Lanciault and Champoux, 2006), so it is not surprising that regions prone to form stem loops (Zhang et al., 2005) and homopolymeric stretches (Baird et al., 2006a) are positively correlated with strand transfer, since they are both associated with pause sites (Klarmann et al., 1993).

The study of *in vivo* recombination has revealed a non-uniform distribution of transfer sites along the genome (Magiorkinis et al., 2003; Renjifo et al., 1999; Takebe et al., 2003; Thomson et al., 2004; Zhang et al., 2005). Inferred breakpoints do not directly reveal mechanistic hotspots of recombination, because these viruses can replicate in hosts and are thus highly selected. In fact, even the selection imposed by multi-cycle *in vitro* recombination assays can change the genomic distribution of strand transfer sites (Baird et al., 2006a). However, we expect strong mechanistic hotspots to leave a signal even after heavy selection. Furthermore, hotspots emerging only post-selection inform on the forces molding the current AIDS epidemic. We have applied a hierarchical model for estimating spatial recombination rates (Minin et al., 2007) to a large dataset of recombinant sequences and identified striking nonuniformities in the distribution of strand transfers along the HIV-1 genome.

## 4.2 Methods

### 4.2.1 Reference Alignment

We downloaded the 2003 reference sequences from HIV-Database (2010), which include full-length sequences representative of each major HIV subtype, circulating recombinant form, and subsubtype. From these reference sequences, consensus sequences for nonrecombinant subtypes (or subsubtypes) A1, A2, B, C, D, F1, F2, G, H, J, and K created by the HIV Database Consensus Tool were realigned using T-Coffee (Notredame et al., 2000), and gaps were trimmed from the alignment ends.

### 4.2.2 Screening for Recombinant Sequences

We downloaded all HIV-1 sequences present in Genbank as of May 17, 2005 at least 400 nucleotides long and discarding patent and other non-natural sequences. Each sequence was aligned to the consensus alignment and the alignment trimmed of gaps at

both ends. To eliminate sequences with very tight epidemiological linkage, we identified groups of sequences with sequential accession numbers, similar lengths (alignments vary less than two nucleotides), and similar sequences (pairwise Hamming distance less than 0.01). One random sequence was selected from each group and all others were not further studied. The resulting data set included $64,603$ HIV-1 sequences.

To reduce the dataset, we analyzed each alignment using cBrother software (Fang et al., 2007), which estimates, via Markov Chain Monte Carlo (MCMC), recombination in a single sequence (Minin et al., 2005). We did not fix the tree relating potential parental subtypes, in order to account for inconsistent relationships among subtypes along the genome (see Fig. 5 of (Anderson et al., 2000)) and ancient recombination among subtypes, which can bias recombination inference if neglected (Fang et al., 2007).

cBrother is inefficient when there are more than six parental sequences and the parental tree is not fixed, so we prepared six separate alignments, each with a subset of five or six parental strains. All pairs of parents are included in at least one of these alignments, so we were able to detect all simple recombinants involving just two parents. For each alignment, we produced an MCMC sample from a short run of length $110,000$, discarding $10,000$ and subsampling every 100. Otherwise, default settings were used. There were $9,819$ simple recombinants detected at this stage. When more than one alignment contained the involved parents, we randomly selected one MCMC sample for future analysis.

We prepared the remanining sequences for more thorough examination, to avoid convergence issues that the intial short runs of MCMC may have suffered. First, we clustered sequences with similar recombinant structure. Clustering on structure serves to eliminate recombinant forms that are descendents of the same recombination event. Their inclusion could falsely inflate recombination signal near their breakpoints. We defined the *profile structure* of a recombinant as the 5' to 3' sequence of parents that are supported with a posterior probability of 0.9 or higher and the posterior medi-

ans of all breakpoints separating adjacent parents. Posterior breakpoint medians and 95% Bayesian credible intervals were computed from the MCMC sample. We clustered sequences with similar structures, i.e. the same parents and co-located breakpoints. Co-located breakpoints were those where either 95% Bayesian credible interval contained the other recombinant's posterior median breakpoint. After randomly selecting one sequence to represent each cluster, there remained $4,073$ sequences. These sequences were analyzed with two independent cBrother runs of length $1,100,000$, discarding $100,000$ and subsampling every $1,000$. We used several convergence diagnostics, including the Gelman-Rubin statistic (Gelman and Rubin, 1992) and all CODA (Best et al., 1995) test statistics with default settings except in the Raftery & Lewis statistic, where convergence was diagnosed for the 0.025 quantile at a precision of $\pm 0.02$ and 90% certainty. Most samples were converged after the initial run, but run lengths were doubled until convergence was achieved for all unconverged samples. After discarding nonrecombinants or complex recombinants identified in the second run, we were left with $2,360$ simple recombinants. Clustering again on profile structure resulted in 544 unique recombinants, representing $1,701$ unique breakpoints.

### 4.2.3 Estimation of Genomic Recombination Rates

Recombinants involving multiple subsubtypes, both B and D and those with subtype K as parent were rarely seen. We, therefore, reduced the number of parents by letting A1 represent subsubtypes A1 and A2, F1 represent F1 and F2, and B represent closely related D (Cornelissen et al., 1997) and removing subtype K. The final, reduced list of parental subtypes was A, B, C, F, G, H and J.

The hierarchical model for combining evidence from multiple recombinants (Minin et al., 2007), implemented in the software BandOfBrothers, seeks to estimate recombination rates $p_s$, the probability of recombination at site $s \in \{1, \ldots, S\}$, along a master alignment of length $S$ containing parental sequences and all recombinants. The model

is developed in terms of the log odds of recombination $\gamma_s = \log \frac{p_s}{1-p_s}$. To constrain the number of free parameters, it places a Gaussian Markov Random Field (Rue and Held, 2005) hyperprior on the parameters $\gamma_s$ and assumes these parameters are correlated for neighboring sites. Evidence that HIV recombination hotspots are not so much spots, rather regions, where strand transfers cluster (Galetto et al., 2006), supports this correlation assumption.

We prepared four data sets: (1) full genome, and three gene-focused datasets (2) *gag*, (3) *pol*, and (4) *env*. We experienced difficulties obtaining a master alignment for the full genome because of the variability in the lengths of the individual recombinants. This master alignment is needed by BandOfBrothers to map breakpoints inferred in individual recombinants to the common indexing system $s$. Our solution was to prepare individual alignments of each recombinant to the full-length reference sequence HXB2 used for numbering genomic positions in HIV-1 (Korber et al., 1998). Then, all breakpoints along the individual recombinants were mapped to their HXB2-relative position. Using this strategy, we lose detailed positional information about breakpoints found within insertions in the recombining parental reference sequences that are not in HXB2. Fortunately, such insertions are rare. Gene-specific master alignments were obtained without difficulty.

The full genome alignment consisted of 544 unique recombinants. The gene-focused alignments consisted of all recombinants with recombinant profile breakpoints falling in the selected gene. Only the gp120 portion of *env* was considered in the *env* dataset, as this is where most studies of experimental recombination have focused. The number of recombinants included in each dataset are 93 in *gag*, 167 in *pol*, and 179 in *env*.

An MCMC sample of the hyperprior parameters $p_s$ was obtained using BandOfBrothers. Briefly, the software starts by sampling from the lower level multiple change point model independently for all recombinants and then alternates between updates of the hyperprior and updates of the lower level change point model conditional on the

hyperprior. Runs used $1,000$ initial samples, then $110,000$ iterations with 10 lower level updates per cycle. We subsampled every 20th iteration after discarding the first $10,000$. All other tuning parameters and hyperparameters were set as suggested (Minin et al., 2007).

### 4.2.4 Analysis of Correlates of Recombination

We considered the relation between recombination rate and two simple features of the genomic primary sequence: GC content and diversity. The posterior distribution of recombination probabilities along the HIV-1 genome is summarized in Fig. 4.1, along with a map of genomic features. Plotted are the posterior median (line) and 95% posterior credible set (shading) for the recombination probability $p_s$ at each site $s$ of the HXB2 genome. We caution that the tested sequences were selected because they were known recombinants, so only *relative* recombination probabilities are meaningful. Furthermore, although we show results for the 5' LTR, the parental reference sequences were not available in this region, so breakpoints could not be inferred there.

For this analysis, we used the 400 random sequences selected for the full genome run. GC content at each site is the proportion of guanine or cytosine nucleotides in the alignment in a window of specified length straddling the site. Window sizes of $20, 50$ and 100 nucleotides were used. A per site estimate of diversity was obtained by computing Shannon's Information entropy at each site $s$

$$H_s = -\sum_{i \in \{A,C,G,U\}} p_{si} \log p_{si}$$

where $p_{si}$ is the proportion of sequences containing nucleotide $i$ at site $s$. $H_s$ values close to zero indicate conserved sites. The R statistics package was used to compute nonparametric Spearman correlations between these numeric summaries of sequences and the recombination rates estimated by BandOfBrothers.

## 4.3 Results

The distribution of breakpoints along the genome is clearly nonuniform. There is a blip in the *gag* gene, though it does not appear significant in the full genome analysis. A pronounced peak is seen in the *pol* gene, specifically in the region encoding the reverse transcriptase. In contrast, the *pol* sequence coding for the RNase and Integrase are strikingly cold. The 3' end of the HIV-1 genome also has higher recombination activity, particularly at both ends of the *env* gene. Overall, *in vivo* recombination breakpoints in sequences sampled from around the world are most dramatically clustered in the 5' portion of the *pol* gene.

We have performed a preliminary analysis to determine whether conservation and GC content are associated with heightened incidence of recombination. Negative lag refers to the sequence feature (entropy or GC content) 3' of the strand transfer site, while positive lag implies 5' sequence features. During minus strand synthesis, 3' features are transcribed before strand transfer occurs. Fig. 4.2(a) shows the correlation of recombination probability $p_s$ with entropy $H_s$ computed at each site $s$. Sequence variability is positively correlated with recombination, especially 400 or fewer nucleotides upstream of the transfer site during minus strand synthesis. Fig. 4.2(b) shows correlation of GC content with recombination rate for different window sizes. There is a negative association of GC content with the site of strand transfer. No correlations are large, especially considering the fact that the data points are not independent along the genome. Effective sample sizes estimated with CODA (Best et al., 1995) suggest the correlations are not statistically significant (data not shown).

We reran the hierarchical model for three local regions of interest: *gag*, *pol*, and *env*. By focusing on recombinant sequences with known breakpoints in the selected genes, we expected to obtain better resolution on the location of recombination activity in these genes. The recombination profile for the *gag* gene (Fig. 4.3(a)) reveals the blip seen in

the full genome analysis is significant. The hotspot appears in p24, which encodes the Capsid protein. As compared to the peak in *gag*, the *pol* peak is more dramatic, with less overlap in credible intervals (Fig. 4.3(b)). Also shown in this plot are the locations of all *in vivo* drug resistance mutations mapping to *pol* (Clark et al., 2005). These mutations cluster in the common drug target, Protease, and the first half of the other common drug target, RT. Interestingly, the hotspot is squarely centered on the cluster of drug resistance mutations associated with RT inhibitors. To investigate whether the hotspot is associated with resistance mutations solely because these mutations represent the only variability in an otherwise conserved gene, we computed average pairwise similarity between dataset sequences. However, similarity peaked just upstream of the recombination hotspot, and the lowest similarity was observed in the 3' of *pol* (data not shown). Finally, the *env* analysis focused on the portion of *env* encoding gp120 (Fig. 4.3(c)) and revealed a concentrated recombination hotspot in the V3 loop.

## 4.4 Discussion

Our analysis of spatial variation in recombination rate combined the data from hundreds of unique HIV-1 recombinants and revealed a distinctly non-uniform distribution of recombination breakpoints along the genome. The overwhelmingly dominant hotspot for recombination was in the reverse transcriptase gene, part of the *pol* open reading frame. Other hot regions include p24 in the *gag* and essentially all of the *env* open reading frame.

The spread of antiretroviral drug treatment around the world has placed the *pol* gene under increasing selective pressure. Given reports that recombination contributes to multi-drug resistance *in vivo* (Charpentier et al., 2006; Nora et al., 2007), the hotspot in *pol* may result from intense selection of random recombinants that happen to combine multiple drug resistance mutations in a single recombinant product. On the other hand,

only 28% of the people needing treatment in low- to middle-income countries were estimated to be receiving antiretroviral treatment by the end of 2006, which itself represents a dramatic increase in the last three years (WHO, 2007). Since subtypes tend to co-circulate mostly in low- to middle-income countries, it remains unclear if selection can be a driving force in producing *pol* inter-subtype recombinants. Of the three previous reports of a hotspot in *pol* (Jetzt et al., 2000; Magiorkinis et al., 2003; Thomson et al., 2004), two (Magiorkinis et al., 2003; Thomson et al., 2004) are also based on *in vivo* recombinants isolated at least three years ago. The third found 5' *pol* to be the hottest site for recombination in a single round infection assay (Jetzt et al., 2000), although the viruses studied excluded the *env* gene. It is unlikely that widespread retroviral drug selection can explain the hotspot in *pol*, but whether the *pol* sequence itself encodes for high strand transfer potential can be assessed in the laboratory, and such experiments are currently underway.

We have previously reported a recombination hotspot in the p24 region of *gag* based on the analysis of 42 AG recombinants (Minin et al., 2007). Here, we analyzed 93 recombinants with all combinations of parents. Although different subtypes may vary in the preferred strand transfer sites (Baird et al., 2006b), we detected a recombination hotspot again in the p24 region of *gag*. In the previous analysis, no attempt was made to screen for circulating recombinant forms (CRFs), but two CRFs are known have an AG breakpoint in the *gag* gene (Kuiken et al., 2010). Here, by clustering recombinants on their recombinant profile structure, we remove both recognized CRFs (Kuiken et al., 2010) and those not yet reported. This discrepancy may explain why the hotspot is less sharply located in our analysis. This portion of *gag* was not a hotspot in an *in vitro* study of recombination, although a *gag* region just downstream was quite recombinogenic (Moumen et al., 2001).

Recombination rates in the *env* gene were generally high (Fig. 4.1). Interestingly, the 5' end and extending into the accessory proteins *vpr*, *tat*, and *vpu* was hottest,

along with the 3' gp41 region. This pattern of recombination activity corroborates the finding that the subtype of *env*, especially gp120, tends to be swapped in both CRFs and unique recombinants (Kuiken et al., 2010; Renjifo et al., 1999; Takebe et al., 2003). Our gp120-focused results revealed a hotspot in V3, which appeared as a mere blip in the full genome. Because the 400 random sequences chosen in the full genome analysis did not include all the 179 sequences of the local analysis, it is possible that the different datasets have produced distinct results. However, the local analysis is unlikely to detect breakpoints at the edges of *env* because there is insufficient upstream or downstream information to infer a topological change (Suchard et al., 2003), so V3 is left as the sole source of supported breakpoints in this region. Also, the specific location of the V3 hotspot could be an artifact of the extreme diversity of this region. Breakpoints may actually locate in the conserved regions C2 or C3, which surround V3 and are confirmed hotspots in experiment (Baird et al., 2006b). The poor alignment plus the effects of the GMRF smoothing hyperprior could draw C2/C3 breakpoints into V3. In summary, although we found evidence of recombination near the heavily reported C2 hotspot, it was by no means the most active spot *in vivo*.

We briefly examined the correlation of sequence conservation and GC content with the inferred full genome recombination rate estimates (Fig. 4.2). While correlations were not strong and probably not significant in our datasets, we discuss here possible reasons for the observed trends. Entropy, a measure of sequence diversity, was positively correlated with recombination rate, in contrast to previous findings of a positive association between *in vivo* recombination and pairwise subtype similarity (Magiorkinis et al., 2003). Similarly, we observed a negative association between average pairwise subtype similarity and recombination rate (data not shown). All the experimental evidence contradicts the idea that recombination is promoted by sequence diversity (Baird et al., 2006b; Gao et al., 2007). Instead, we hypothesize that the association results because breakpoints are easier to detect and localize in regions of higher diversity. In other words,

the results may reflect the fact that the underlying methodology (and any method) will increasingly fail to detect recombination as sequence conservation increases (Posada and Crandall, 2001). In contrast, we observed that GC content was negatively correlated with recombination rate. This finding is odd considering that RT pausing occurs on GC runs during minus strand synthesis (Klarmann et al., 1993). On the other hand, AT excess may facilitate the dissociation and reassociation required for strand transfer, independent of other predisposing features.

We have revealed that the *in vivo* pattern of recombination breakpoints along the HIV-1 genome is highly non-uniform. We restricted our analysis to simple recombinants involving at most two parental sequences, and we did not consider CRFs as parents, which would allow detection of second generation recombinants (Toni et al., 2005). Inclusion of more recombinants in the future will allow better resolution of spatial recombination variation. While we tried to control for the presence of repeatedly sampled recombinants by clustering on recombinant structure, we cannot guarantee that every one of the $1,701$ breakpoints in our final dataset represents a unique recombination event. However, if a breakpoint appears repeatedly in multiple contexts, then it is likely selected. Ultimately, separation of mechanism and selection requires additional experimentation and theoretical models. These results highlight more interesting regions to target in such future study.

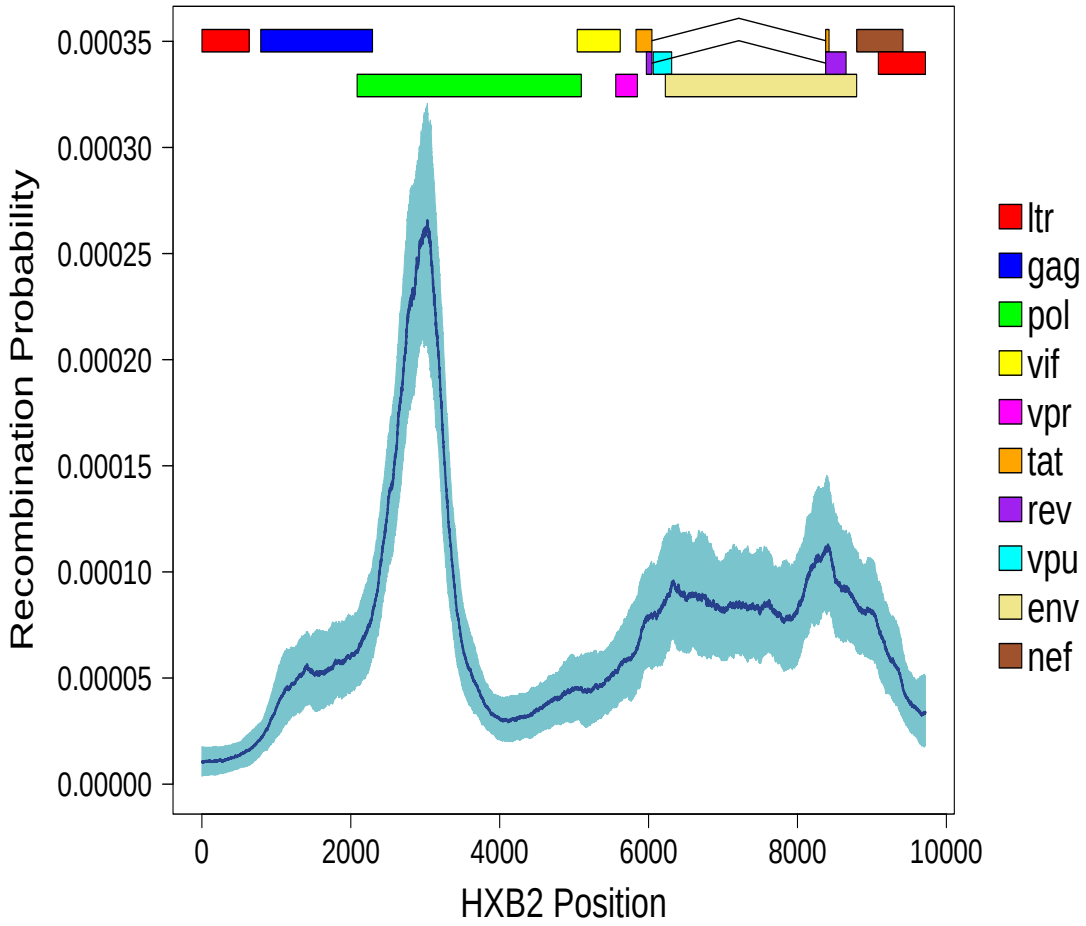
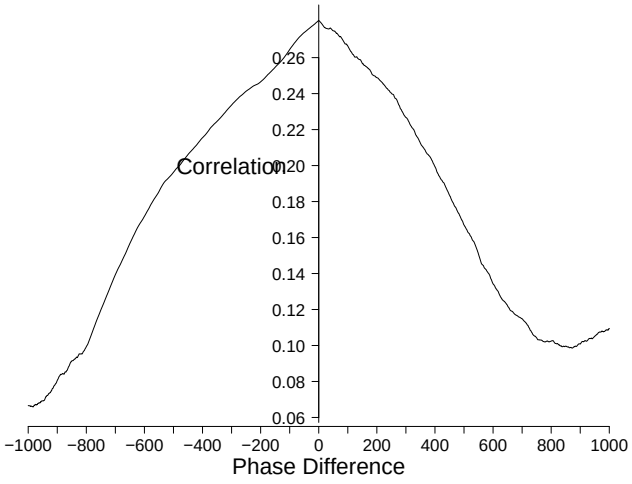
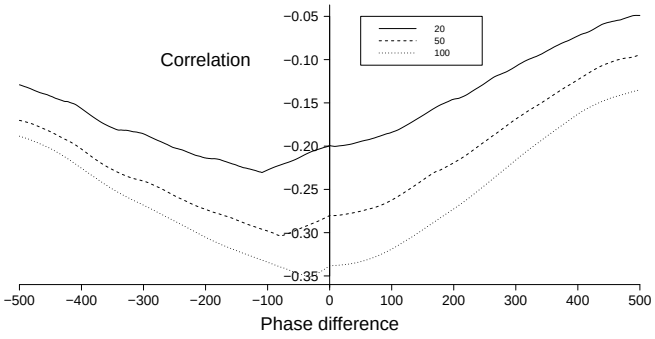



Figure 4.1 Spatial recombination profile of the full HIV-1 genome.




(a)




(b)

Figure 4.2 Lag correlation of recombination and (a) entropy and (b) GC content for window sizes 20,50, and 100.

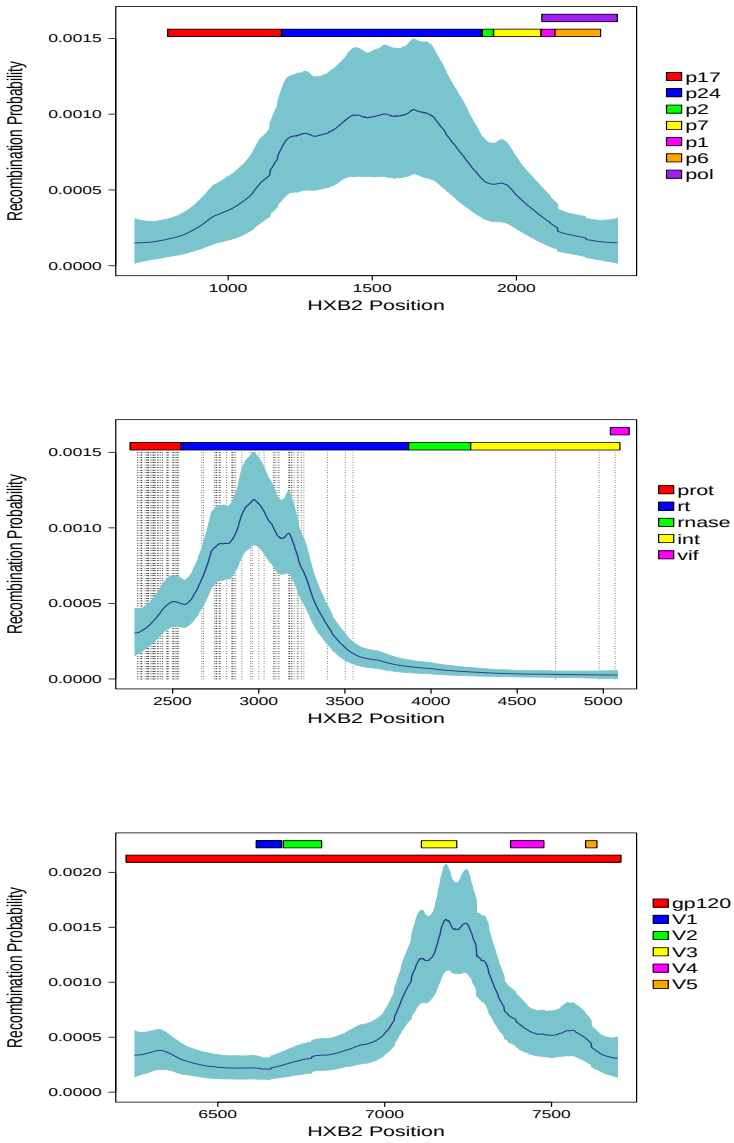



Figure 4.3 Spatial recombination profile for (a) *gag* (b) *pol* (c) *env*

# CHAPTER 5. REVIEW II

This chapter reviews, in detail, the statistical theory and techniques used in the following chapters. The final section serves as a small introduction to the model presented in the next chapter.

## 5.1 Markov Chain Monte Carlo : A Brief Primer

The goal of MCMC procedures is to sample from a target distribution. Instead of sampling directly from the target, we sample from a Markov chain constructed to have the target distribution as its equilibrium distribution. Given enough samples and appropriate conditions on the chain, theory guarantees that the dependent samples will be drawn from the target distribution (Tierney, 1994).

### 5.1.1 Metropolis-Hastings Algorithm

It is simplest to construct a *reversible* Markov chain, where the transition probabilities $p(\cdot;\cdot)$ and equilibrium distribution $\pi(\cdot)$ satisfy detail balance,

$$\pi(\theta)p(\theta;\theta^*) = \pi(\theta^*)p(\theta^*;\theta). \tag{5.1}$$

Here, $\theta$ is a vector of model parameters. Applying principles of rejection sampling (Robert and Casella, 2005), this transition kernel can be expressed as

$$p(\theta;\theta^*) = q(\theta;\theta^*)\alpha(\theta;\theta^*), \tag{5.2}$$

for any arbitrary proposal distribution $q(\theta; \theta^*)$, proposing $\theta^*$ given current state $\theta$, and an acceptance probability $\alpha(\theta; \theta^*)$, insuring the stationary distribution $\pi(\theta)$ is the desired target distribution. Hastings (1970) shows acceptance probability

$$\alpha(\theta; \theta^*) = \min \left[ 1, \frac{\pi(\theta)}{\pi(\theta^*)} \frac{q(\theta^*; \theta)}{q(\theta; \theta^*)} \right] \tag{5.3}$$

yields the desired equilibrium distribution $\pi(\theta)$. To guarantee detail balance (5.1), any transition from parameters $\theta$ to $\theta^*$ with $q(\theta; \theta^*) > 0$ must be reversible, so $q(\theta^*; \theta) > 0$, when the parameters are valid, i.e. $\pi(\theta), \pi(\theta^*) > 0$ (Tierney, 1994). The resulting Metropolis-Hastings MCMC algorithm (Chib and Greenberg, 1995) samples $\theta_1, \theta_2, \ldots$ from any target distribution $\pi(\theta)$ in the following procedure.

1. Start with some initial value $\theta_0 = \theta^0$.

2. Given the current iterate $\theta_t$, repeat until convergence:

   (a) Draw $\theta^* \sim q(\theta_t; \theta^*)$.

   (b) Draw $u \sim \text{Unif}(0, 1)$.

   (c) Compute $\alpha(\theta_t; \theta^*)$.

   (d) If $\alpha(\theta_t; \theta^*) \geq u$, accept the proposal and set $\theta_{t+1} = \theta^*$.

### 5.1.2 Gibbs Sampling

Gibbs Sampling (Resnik and Hardisty, 2009) uses full conditional distributions of a multivariate parameter to sample from their joint posterior. Suppose that we wish to generate samples from $\pi(\boldsymbol{\theta})$ where now $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_p)$. Without loss of generality, let us consider a case where $p = 2$. The full conditionals of interest are $\pi(\theta_1|\theta_2)$ and $\pi(\theta_2|\theta_1)$.

The Gibbs sampling algorithm prescribes the following procedure:

1. Start with some initial value $\boldsymbol{\theta^0} = (\theta_1^0, \theta_2^0)$.

2. Given the current iterate $\boldsymbol{\theta^t}$, repeat until convergence:

   (a) Draw $\theta_1^* \sim \pi(\theta_1|\theta_2 = \theta_2^t)$

   (b) Draw $\theta_2^* \sim \pi(\theta_2|\theta_1 = \theta_1^*)$

Note that this scheme defines a reversible Markov chain that satisfies the detailed balance condition and has equilibrium distribution $\pi(\boldsymbol{\theta})$. Each iteration of the Gibbs sampler may be viewed as $p$ iterations of the Metropolis-Hastings sampler (Gelman et al., 2004). However, the proposal distribution changes every iteration.

$$q(\boldsymbol{\theta}^*, \mid \boldsymbol{\theta}^t) = \begin{cases} p(\theta_1^* \mid \theta_2^t, \boldsymbol{y}) \text{ if } \theta_2^* = \theta_2^t \\ \quad 0 \text{ otherwise.} \end{cases}$$

Consider the acceptance probability after a proposal for $\theta_1$ has been drawn. Here, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^t)$ and the current iterate $\boldsymbol{\theta}^t = (\theta_1^t, \theta_2^t)$.

$$\begin{aligned} \alpha(\boldsymbol{\theta}^*; \boldsymbol{\theta}^t) &= \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \frac{q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^i)} \\ &= \frac{p(\theta_1^*, \theta_2 \mid \boldsymbol{y})}{p(\theta_1, \theta_2 \mid \boldsymbol{y})} \frac{p(\theta_1 \mid \theta_2, \boldsymbol{y})}{p(\theta_1^* \mid \theta_2, \boldsymbol{y})} \\ &= \frac{p(\theta_1^*, \theta_2 \mid \boldsymbol{y}) p(\theta_1, \theta_2 \mid \boldsymbol{y}) p(\theta_2 \mid \boldsymbol{y})}{p(\theta_1, \theta_2 \mid \boldsymbol{y}) p(\theta_1^*, \theta_2 \mid \boldsymbol{y}) p(\theta_2 \mid \boldsymbol{y})} \\ &= 1 \end{aligned} \tag{5.4}$$

Thus every proposal is accepted.

The two algorithms described above may also be combined. One such method is the Metropolis-within-Gibbs algorithm. Here, each parameter is individually updated following Gibbs sampling. However, in cases where the full conditionals are intractable, proposal distributions may be used and individual Metropolis-Hastings acceptance/rejection employed to obtain updated parameters.

### 5.1.3 Multiple Changepoint Models and rjMCMC

Many times we are faced with a problem where the dimension of $\theta$ is unknown and must be estimated. To be explicit, suppose that we have a set of responses $\boldsymbol{y}$ that arise from one of a countable number of possible models $M_0, M_1 \ldots M_K$. Each model $M_k, k = 0, \ldots, K$, has a parameter vector $\theta_k$ and the dimensionality of $\theta_k$ may vary with model $k$. We are specifically concerned with multiple changepoint models, where an ordered response vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ is hypothesized to be partitioned at $k$ changepoints $s_1, \ldots, s_k \in \{2, \ldots, n\}$. Between changepoints $i$ and $i+1$, the data $y_{s_i}, y_{s_i+1}, \ldots, y_{s_{i+1}-1}$ are iid realizations of some distribution $f(y; \mu_i)$, say a normal with known variance $\sigma^2$. Model $M_0$ corresponds to no changepoints, so all $(y_1, \ldots, y_n) \overset{\text{iid}}{\sim} N(\mu_0, \sigma^2)$ and $\theta_0 = \mu_0$. Model $M_1$ has one changepoint, resulting in a parameter space with two means and a change point location, $\boldsymbol{\theta}_1 = (\mu_0, \mu_1, s_1)$. Similarly $\boldsymbol{\theta}_2 = (\mu_0, \mu_1, \mu_2, s_1, s_2)$ has another two dimensions and so on. The goal is to infer the model index as well as the associated parameters using the MCMC scheme and due to changing dimensions of the parameter space, the rjMCMC scheme fits our needs very well.

#### 5.1.3.1 Model

We describe a simple change point model. The purpose is to use simulation and a simple model to illustrate the same approach we will use to model recombination-induced change points in sequence data.

If parameters of model $M_k$ are $\boldsymbol{\theta}_k = (\boldsymbol{\mu}, \boldsymbol{s})$, our goal is to sample from the joint posterior

$$\pi(k, \boldsymbol{\theta}_k \mid \boldsymbol{y}) \propto L(\boldsymbol{y} \mid k, \boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k \mid k)P(k). \tag{5.5}$$

Conditional on the change points and means, the data are independent draws from

normal distributions, so our likelihood is

$$L(\boldsymbol{y} \mid k, \boldsymbol{\theta}_k) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma}\left(y_i - \mu(i)\right)^2\right] \tag{5.6}$$

where $\mu(i) = \mu_j$ for all $i$ belonging to partition $j$.

In the Bayesian context, we must propose a prior for all model parameters. We assume the number of change points has a truncated Poisson prior

$$P(k) \propto \frac{e^{-\lambda}\lambda^k}{k!}, \qquad 0 \leq k < n,$$

where the constant of proportionality is the Poisson probability $P(k < n)$. Conditional on the number of change points, the locations and heights are independent. The locations are order statistics of discrete draws without replacement on $\{2, \ldots, n\}$,

$$\pi(\boldsymbol{s} \mid k) = \frac{k!}{(n-1)\cdots(n-k)}.$$

The normal means $\mu_j$ are i.i.d. normal $N(5, 1)$, specifically

$$\pi(\mu_j) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\mu_j - 5\right)^2\right].$$

#### 5.1.3.2 Reversible Jump MCMC (rjMCMC)

In order to use MCMC techniques to make inference on problems with variable dimension, Green (1995) suggested a reversible jump MCMC (rjMCMC) algorithm. As in many MCMC procedures, rjMCMC alternates among different move types in order to explore the parameter space. In our case, we define three possible moves types:

1. Stay (S): update parameters without changing the dimension,

2. Birth (B): addition of a change point, and

3. Death (D): deletion of a change point.

The Birth and Death moves are trans-dimensional move types that require special proposal distributions, with more involved calculation of proposal ratios, discussed in later sections.

At each iteration of the MCMC algorithm, i.e. each transition, move S is attempted with probability $e_k$, B with probability $b_k$, and D with probability $d_k$, such that $e_k + b_k + d_k = 1$. These move probabilities depend only on the current number of changepoints $k$. Since no more than $n-1$ change points are allowed, we have $b_{n-1} = 0$. Otherwise, we set

$$\begin{aligned} b_k &= c \min\left(1, \frac{P(k+1)}{P(k)}\right), \\ d_k &= c \min\left(1, \frac{P(k-1)}{P(k)}\right) \text{ and} \\ e_k &= 1 - b_k - d_k, \end{aligned}$$

where the constant $c$ is chosen to be as large as possible while maintaining $e_k > 0.1$ for all possible values of $k$. Since $b_k + d_k$ is maximized at $k = \lambda$ , it is not hard to see that setting

$$c = \frac{0.9(\lambda+1)}{2\lambda+1}$$

guarantees the conditions.

#### 5.1.3.3 Birth Move

The key to trans-dimensional moves in rjMCMC is a bijective map $f : (\boldsymbol{\theta}_k, \boldsymbol{\xi}) \leftrightarrow (\boldsymbol{\theta}_{k+p})$ between the lower dimensional parameter vector $\boldsymbol{\theta}_k$ to the higher dimensional parameter vector $\boldsymbol{\theta}_{k+p}$. To define the bijection, $\boldsymbol{\theta}_k$ must be supplemented with vector $\boldsymbol{\xi}$ of length $p$. The acceptance ratio for this setup is (Green, 1995)

$$\alpha(\boldsymbol{\theta}_k; \boldsymbol{\theta}^*_{k+p}) = \min\left[1, \frac{\pi(\theta)}{\pi(\theta^*)} \frac{q(\boldsymbol{\theta}^*_{k+p}; \boldsymbol{\theta}_k)}{q(\boldsymbol{\theta}_k; \boldsymbol{\theta}^*_{k+p})} \left|\frac{\partial \boldsymbol{\theta}^*_{k+p}}{\partial \boldsymbol{\theta}_k}\right|\right], \tag{5.7}$$

where $q(\boldsymbol{\theta}^*_{k+p}; \boldsymbol{\theta}_k)$ is the proposal density for moving $\theta_{k+p} \to \theta_k$, $q(\boldsymbol{\theta}_k; \boldsymbol{\theta}^*_{k+p})$ is the proposal density for moving $\theta_k \to \theta_{k+p}$, and $\left|\frac{\partial \boldsymbol{\theta}^*_{k+p}}{\partial \boldsymbol{\theta}_k}\right|$ is the Jacobian for the bijection.

We demonstrate below how to choose $\boldsymbol{\xi}$ and compute the acceptance ratios $\alpha(\cdot;\cdot)$ for our simple change point model.

Birth moves jump between parameter spaces $\theta_k$ and $\theta_{k+1}$, specifically two states

$$\boldsymbol{\theta}_k = (\mu_0, \ldots, \mu_{k+1}, s_1, \ldots, s_k) \in \theta_k, \text{ and}$$
$$\boldsymbol{\theta}_{k+1} = (\mu_0^*, \ldots, \mu_j^*, \mu_{j+1}^*, \ldots, \mu_{k+2}^*, s_1^*, \ldots, s_{j+1}^*, \ldots, s_{k+1}^*) \in \theta_{k+1}.$$

Thus, given $k$ change points, a birth move increases the dimension from $2k+1$ to $2k+3$. To match dimension, we must define two variables $\boldsymbol{\xi} = (s, u)$ to supplement parameter vector $\boldsymbol{\theta}_k$. As usual, MCMC leaves a lot of choices up to the user. We choose to draw $s$ uniformly from the $n-k-1$ locations that are not already changepoints to represent the new change point $s_{j+1}^*$. In addition $u$ is a standard uniform random deviate used to generate the new means $(\mu_j^*, \mu_{j+1}^*)$ from supplemented old mean $(u, \mu_j)$ according to the following bijective map

$$u = \frac{\mu_j^*}{\mu_j^* + \mu_{j+1}^*} \tag{5.8}$$
$$\mu_j = \frac{s_{j+2}^* - s_{j+1}^*}{s_{j+2}^* - s_j^*}\mu_{j+1}^* + \frac{s_{j+1}^* - s_j^*}{s_{j+2}^* - s_j^*}\mu_j^*. \tag{5.9}$$

Uniform deviate $u$ determines the variability in new means $\mu_j^*$ and $\mu_{j+1}^*$ and $\mu_j$ is mapped as the weighted average of $\mu_j^*$ and $\mu_{j+1}^*$. With $s = s_{j+1}^*$, all other parameter are mapped one-to-one in the full bijection.

From the above description, the birth proposal is

$$q(\boldsymbol{\theta}_k; \boldsymbol{\theta}_{k+p}^*) = b_k q(s, u) = \frac{b_k}{n-k-1},$$

where we recognize that after choosing to attempt a birth move with probability $b_k$ and generating the independent random variables $s, u$ with probability $\frac{1}{n-k-1}$, the map is deterministic. For the reversed death distribution, we have

$$q(\boldsymbol{\theta}_{k+1}^*; \boldsymbol{\theta}_k) = \frac{d_{k+1}}{k+1},$$

where we choose to lose a change point with probability $d_{k+1}$, select the change point with uniform probability $\frac{1}{k+1}$, and progress deterministically from there. The important part of the Jacobian for this bijection scheme is

$$\left|\begin{array}{cc} \frac{\partial \mu_j}{\partial \mu_j^*} & \frac{\partial \mu_j}{\partial \mu_{j+1}^*} \\ \frac{\partial u}{\partial \mu_j^*} & \frac{\partial u}{\partial \mu_{j+1}^*} \end{array}\right| = \left|\begin{array}{cc} \frac{s_{j+2}^*-s_{j+1}^*}{s_{j+2}^*-s_j^*} & \frac{s_{j+1}^*-s_j^*}{s_{j+2}^*-s_j^*} \\ \frac{\mu_{j+1}^*}{(\mu_j^*+\mu_{j+1}^*)^2} & \frac{-\mu_j^*}{(\mu_j^*+\mu_{j+1}^*)^2} \end{array}\right| = \left|\frac{(s_{j+2}^*-s_{j+1}^*)\mu_j^* + (s_{j+1}^*-s_j^*)\mu_{j+1}^*}{(s_{j+2}^*-s_j^*)(\mu_j^*+\mu_{j+1}^*)^2}\right|. \tag{5.10}$$

Putting it all together, we have

$$\begin{aligned} \alpha(\boldsymbol{\theta}_k;\boldsymbol{\theta}_{k+1}^*) &= \min\left\{1, \frac{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_{k+1}^*)}{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_k)} \times \frac{\pi(\mu_j^*)\pi(\mu_{j+1}^*)}{\pi(\mu_j)} \times \frac{k+1}{n-k-2}\right. \\ &\quad \left. \times \frac{\lambda}{k+1} \times \frac{d_{k+1}(n-k-1)}{b_k(k+1)} \times \left|\frac{(s_{j+2}^*-s_{j+1}^*)\mu_j^* + (s_{j+1}^*-s_j^*)\mu_{j+1}^*}{(s_{j+2}^*-s_j^*)(\mu_j^*+\mu_{j+1}^*)^2}\right|^{-1}\right\} \end{aligned}$$

where $\mathcal{L}(;)$ denotes the likelihood and $\pi(\mu)$ denotes the prior distribution on $\mu$.

#### 5.1.3.4 Death Move

The acceptance ratio in the death move is the inverse of the acceptance ratio for the birth move from state $k-1$ to state $k$.

$$\begin{aligned} &\alpha(\boldsymbol{\theta}_k;\boldsymbol{\theta}_{k-1}^*) = \\ &\min\left\{1, \frac{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_{k-1}^*)}{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_k)} \times \frac{\pi(\mu_j^*)}{\pi(\mu_j)\pi(\mu_{j+1}) - 5)}\right. \\ &\left. \times \frac{n-k-1}{k} \times \frac{k}{\lambda} \times \frac{b_{k-1}(k)}{d_k(n-k)} \times \left|\frac{(s_{j+2}-s_{j+1})\mu_j + (s_{j+1}-s_j)\mu_{j+1}}{(s_{j+2}-s_j)(\mu_j+\mu_{j+1})^2}\right|\right\} \end{aligned}$$

#### 5.1.3.5 Stay Move

If neither the birth nor death move is chosen, then we update each $s_j$ and $\mu_j$ in a Metropolis-within-Gibbs step (Gilks et al., 1995).

- Define $s_0 = 0$ and $s_{k+1} = n$, then for each $s_j \in \boldsymbol{s}$
  - Propose a new $s_{j+1}^*$ uniformly in the interval $(s_{j-1}, s_{j+1})$.

  - The acceptance probability is

$$\min\left\{1, \frac{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_k^*)}{\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}_k)}\right\} \tag{5.11}$$

- For each $\mu_j \in \boldsymbol{\mu}$
  - Propose a new $\mu_j^* \sim \pi(\mu)$.
  - The acceptance ratio is

$$\min\left\{1, \frac{L(\boldsymbol{y};\boldsymbol{\theta}_k)}{L(\boldsymbol{y};\boldsymbol{\theta}_k)} \times \frac{\pi(\mu_j^*)}{\pi(\mu_j)}\right\}. \tag{5.12}$$

## 5.2 Gaussian Markov Random Fields

Consider a random normal vector $x = x_1, x_2 \dots x_s$ which has, conditionally independent components

$$x_i \perp x_j | x_{-ij} \tag{5.13}$$

i.e. $x_i$ and $x_j$ are independent conditional on all other components of $x$. Using a graph $G = (V, E)$ to represent the structure of this conditional independence, we have $V$ as the set of vertices, one for each $x_i$ and $E$ as the set of edges, connecting any two vertices that are dependent. $x$ is now a GMRF with respect to $G$ with mean $\mu$ and precision matrix $Q$ if and only if

$$f(\boldsymbol{x}) \sim \mathcal{N}(\mu, Q^{-1}). \tag{5.14}$$

It is easier to parametrize the GMRF in terms of its precision matrix since $Q_{ij} = 0$ when $i \perp j$ and hence it is often a sparse matrix and easier to work with than a dense covariance matrix $\Sigma$.

In some applications, the precision matrix may not be full-rank. This could be due to the structure of $G$ or more commonly due to linear constraints imposed on the GMRF. GMRFs with sub-rank precision matrices are called *Improper* GMRFs.

Of special interest are intrinsic GMRFs (IGMRF) that are improper GMRFs with a precision matrix of rank $n-1$ and $Q1 = 0$ meaning that

$$\sum_j Q_{ij} = 0 \tag{5.15}$$

i.e. rows in the matrix sum to zero. Rue and Held (2005) illustrate the construction of an IGMRF of first order on the line meaning that every node only has its two immediately adjacent node, one on either side, in its neighborhood. The joint density of such a vector $\boldsymbol{x}$ is given by

$$f(\boldsymbol{x}) \propto \omega^{(n-1)/2} \exp\left(-\frac{\omega}{2}\sum_{i=1}^{i=n}(x_i - x_{i-1})^2\right) \tag{5.16}$$

where $\omega$ is the multiplier for the precision matrix which has structure,

$$Q = \omega \times \begin{pmatrix} -1 & 1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & -1 & 2 & -1 & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 & \end{pmatrix} \tag{5.17}$$

This prior, therefore penalizes large differences in recombination probabilities of adjacent sites. The precision matrix Q as in Eq. 5.17. It follows that $Pr(\boldsymbol{x}|\omega) \sim \mathcal{N}(0, Q^{-1})$ and $\boldsymbol{Q}$ is non-full rank. Sun et al. (1999) showed that despite the improper prior, most likelihood functions will still result in a proper posterior. The few that do not, notably, include the binomial and Poisson distributions. In cases where our model requires use of these distributions, we replace $Q$ in Eq. 5.16 with a proper approximation such that $Pr(\boldsymbol{x}|\omega) \sim N(0, \tilde{Q}^{-1})$ where $\tilde{Q} = Q + \epsilon I$ and $\epsilon$ is a small positive constant and $\boldsymbol{I}_{N\times N}$ is an identity matrix.

| **Algorithm for sampling from $\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$** |
|---|
| 1: Compute Cholesky Factorization, $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ |
| 2: Solve $\boldsymbol{L}\boldsymbol{w} = \boldsymbol{b}$ |
| 3: Solve $\boldsymbol{L}^T\boldsymbol{\mu} = \boldsymbol{w}$ |
| 4: Sample $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ |
| 5: Solve $\boldsymbol{L}^T\boldsymbol{\nu} = \boldsymbol{z}$ |
| 6: Compute $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\nu}$ |
| 7: Return $\boldsymbol{x}$ |

Table 5.1 Algorithm for simulation from a GMRF

### 5.2.1 Sampling from a GMRF

**Canonical Parameterization of a GMRF** For any normal random vector $\boldsymbol{X}$, the canonical parameterization can be obtained as follows:

$$\begin{aligned} X &\sim N(\mu, \Sigma) \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X} + \boldsymbol{b}^T\boldsymbol{X}\right) \\ &\sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q}) \end{aligned} \tag{5.18}$$

where $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{\mu}$ and thus the mean of the original normal is $\boldsymbol{\mu} = b\boldsymbol{Q}^{-1}$ and variance is $\boldsymbol{Q}^{-1}$. Such parameterization may be extended to any exponential family distribution.

Thus, a GMRF $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$, precision matrix $\boldsymbol{Q}$ and canonical parameter $\boldsymbol{b} = \boldsymbol{Q}^{-1}\boldsymbol{\mu}$ is expressed in canonical parameterization as :

$$\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$$

In order to simulate from this GMRF, Rue and Held (2005) prescribe the following algorithm

## 5.3 Hierarchical GMRF Models

GMRFs find a wide variety of applications, the most useful of which involve hierarchical models with GMRF hyperpriors. Suppose that we observe some data $\boldsymbol{y}$ that are dependant on a GMRF $\boldsymbol{x}$. We assume that members of $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{x}$. Let $\boldsymbol{\theta}$ denote the hyperparameters that specify the GMRF. The hierarchical setup now looks like this:

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \\ \boldsymbol{x} &\sim \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ y_i &\overset{\text{iid}}{\sim} \pi(y_i \mid x_i), i = 1, 2 \dots n \end{aligned} \tag{5.19}$$

The posterior distribution is then

$$\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i \mid x_i) \tag{5.20}$$

As a simple example, suppose that at the *lower level* or *first stage* of the hierarchical model we observed data $\boldsymbol{y} = y_1, y_2 \dots y_n$ that are counts of incidence of a disease in $n$ counties of a state. For this data we may assume a Poisson model with mean $\boldsymbol{c}e^{\boldsymbol{x}}$ i.e.

$$y_i \sim \mathcal{P}(c_i \exp(x_i))$$

where $\boldsymbol{c}$ represents some known constant, the population in each county for example. Let $\boldsymbol{x}$ represent the probability or incidence of the disease. The GMRF prior is placed on $\boldsymbol{x}$ with precision $\omega$ and a $\boldsymbol{Q}$ matrix that reflects the actual spatial neighborhood of each county (Besag et al., 1991; Rue and Held, 2005).

### 5.3.1 Inference Framework

Statistical inference of hierarchical models involving GMRF hyperpriors is done using Markov Chain Monte Carlo algorithms. The choice of the algorithm used depends on the lower level response model.

#### 5.3.1.1 Inference on Normal response models

Recall that in a hierarchical model setup, we wish to sample from the joint posterior distribution (5.20). Let us consider models where the response can be modeled using a Gaussian distribution i.e $\pi(y_i|x_i) \sim \mathcal{N}(x_i, \sigma_i^2)$. We can now express the combined likelihood as a multivariate normal distribution

$$\pi(\boldsymbol{y}|\boldsymbol{x}) \sim \text{MVN}(\boldsymbol{x}, M) \tag{5.21}$$

where $M$ denotes the precision matrix, which is this case is a $N \times N$ matrix with entries $1/\sigma_i^2$ along the diagonal and zero elsewhere. Substituting in (5.20), the full conditional is now

$$\begin{aligned}
\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) &\propto exp(-\frac{1}{2}(\boldsymbol{x})^T\boldsymbol{Q}(\boldsymbol{x}) + -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T\boldsymbol{M}(\boldsymbol{y}-\boldsymbol{x})) \\
&= exp(-\frac{1}{2}(\boldsymbol{x})^T\boldsymbol{Q}(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{M}\boldsymbol{y} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{M}\boldsymbol{x}) \\
&\propto exp(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{Q}+\boldsymbol{M})\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{M}\boldsymbol{x}) \\
&\sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q}+\boldsymbol{M}) \text{ where } \boldsymbol{b} = \boldsymbol{M}\boldsymbol{y} \\
&\sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q} + diag(\boldsymbol{c}))
\end{aligned} \tag{5.22}$$

Expressed in this form, sampling proceeds using the algorithm outlined in Table 5.1. Since we are drawing directly from a GMRF, this may be viewed as a Gibbs step.

#### 5.3.1.2 Inference on Non-normal response models

Often we may be interested in modeling non-normal data through hierarchical GMRF priors. In these cases, the likelihood is non-normal leading to a joint posterior that, unlike the normal response model, does not retain its Gaussian properties. A Metropolis-Hastings step is used to sample from such joint posteriors. To obtain a reasonable proposal distribution, we build a *GMRF approximation* of the joint posterior by replacing the likelihood with its second-order Taylor series expansion (Rue and Held, 2005).

**Second - order Taylor expansion** Suppose that we want to approximate some function $f(x)$ using a quadratic Taylor expansion. Upon expansion around some point $x_0$, $f(x)$ can be expressed as

$$\begin{aligned} f(x) &\approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 \\ &= a + bx - \frac{1}{2}cx^2 \end{aligned}$$

where,

$$\begin{aligned} a &= f(x_0) + f'(x_0)x_0 + \frac{1}{2}f''(x_0)x_0^2 \\ b &= f'(x_0) - f''(x_0)x_0 \\ c &= -f''(x_0) \end{aligned} \tag{5.23}$$

Applying this expansion to the likelihood term in (5.20),

$$\begin{aligned} \tilde{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu}) + \sum_i (a_i + b_i x_i - \frac{1}{2}c_i x_i^2)\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T(\boldsymbol{Q} + diag(\boldsymbol{c}))(\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{b})^T\boldsymbol{x}\right) \\ &\sim \mathcal{N}_C(\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{Q} + diag(\boldsymbol{c})) \end{aligned} \tag{5.24}$$

Sampling from this follows as before. However, we have an additional Metropolis-Hastings acceptance step here to accept or reject the proposal drawn from the GMRF approximation.

### 5.3.2 Blocking Strategies for Hierarchical GMRF Models

In a hierarchical setup with a GMRF hyperprior, precision $\omega$ is a smoothing parameter and may be fixed at an empirically determined value or sampled from the joint posterior. In the later case, we may place a $\Gamma(a, b)$ hyperprior on it. When $\omega$ is estimated, a block update strategy is employed (Rue and Held, 2005), i.e. the proposal $(\omega^*, \boldsymbol{x}^*)$ is accepted/rejected jointly. This is denoted as the *one-block algorithm*. The

steps involved to are to sample

$$\begin{aligned} \omega^* &\sim q(\omega^* \mid \omega) \\ \boldsymbol{x}^* &\sim \pi(\boldsymbol{x} \mid \boldsymbol{\theta}^*, \boldsymbol{y}) \end{aligned} \tag{5.25}$$

Further, Rue and Held (2005) suggest the use of a symmetric proposal distribution for $\omega$ stating that the $\omega$ chain thus generated is in fact sampling from the posterior marginal $\pi(\omega \mid \boldsymbol{y})$.

### 5.3.3 Extensions of the Hierarchical GMRF Model

Gaussian Markov Random Fields or Intrinsic CAR(1) priors have historically found wide applications in spatial modeling. In a seminal paper, Besag et al. (1991) proposed the application of standard Bayesian image restoration models to disease models. To illustrate this model, consider the notation and example that they use. Suppose that we want to model the risk of a disease over an area made of contiguous subunits or "zones" (for example counties in a state, states in a country). Let $\eta_i$ denote the log relative risk in zone $i \in (1, 2, \ldots n)$ and $y_i$ be the corresponding observed number of cases.

We now assume that for a non-contagious and rare disease (cancer for example),

$$p(y_i|x_i) \sim \text{Poisson}(c_i e^{x_i}) \tag{5.26}$$

where $c_i$ is the expected number of cases in zone $i$ assuming constant risk. In order to jointly model the effect of fixed covariates and error terms, we express $\boldsymbol{x}$ as

$$x = \alpha + u + v$$

where $\alpha$ relates to some known/ fixed covariates akin to the intercept term, $u$ to a spatially structured error term and $v$ to random noise.

**Priors** $u$ and $v$ are assumed to be independent and $v$ $\mathcal{N}(0, \lambda^2)$. On $u$ we put a GMRF prior such that

$$p(u|\omega) \approx \omega^{(n-1)/2} \exp\left(\frac{-\omega}{2}\sigma(u_j - u_i)^2\right) \tag{5.27}$$

$$p(v|\tau) \approx \tau^{n/2} \exp\left(\frac{-\tau}{2}\sigma v_i^2\right) \tag{5.28}$$

Note that expressing (5.28) as

$$p(x|u, \tau) \approx \tau^{n/2} \exp\left(-\frac{\tau}{2}\sigma(x_i - u_i - \alpha)^2\right) \tag{5.29}$$

yields a conditional Gaussian distribution for the $(x_i \mid u_i, \tau)$ with mean $(u_i + \alpha)$ and precision $\tau$.

#### 5.3.3.1 Ecological regression model

The BYM model (Besag et al., 1991) can be extended further to include area level covariates $Z_i$ by expressing $x$ as:

$$x_i = \alpha + u_i + v_i + Z_i\beta \tag{5.30}$$

where $\beta$ is the coefficient of association and $Z_i$ pertains to the model matrix for observation $i$. A uniform prior is applied to $\beta$. This model is especially elegant in that it enables one to not only smooth empirical data over "geographical" areas but also to find links or associations of this data with covariates. Bernardinelli et al. (1997) extended the ecological regression model to allow for covariates with errors, i.e. random covariates. This is common is cases where $Z_i$ can not be observed directly. Suppose that we observe some $w_i$ instead, the simplest approach is to use these to get estimates of $Z_i$, i.e. $\hat{Z}_i$, and use these for calculation. However, this approach relies heavily on $w_i$ being a good approximation of $Z_i$. When this is not the case this can yield underestimated $\beta$ values or overestimated precision measures. Additionally, if we have reason to believe that the $Z_i$ themselves have a spatially correlated structure, then we can obtain better estimates

overall by applying a spatial smoothing prior to $Z_i$ as well that is parameterized by its own precision parameter $\omega_Z$.

## 5.4 Quantifying Association of Recombination Probabilities with Covariates

Minin et al. (2007) presented a pioneering model that enables inference of population level recombination probabilities through the use of a GMRF hyperprior that is able to combine information of recombination events inferred from individual variants (referred hereafter as the Minin model). In Chapter 4, I presented a large-scale application of their model to full length genomic data in order to gain insights into occurrence of recombination hotspots in the HIV genome. In Chapter 6 we present an extension of this model that enables inference of coefficients of association of interesting covariates. Further, we address certain data restrictions imposed by the Minin model. We present these here briefly. The following chapter provides details.

**Mapping to HXB2-relative positions:** The Minin model requires the use of a multiple sequence alignment of all sequences in a dataset. The need for such an alignment precludes the analysis of large datasets. We propose mapping to HXB2-relative positions instead creating a full multiple sequence alignment. While we lose some data pertaining to gaps in HXB2, this is a reasonable trade-off for the ability to include a large number of sequences in the analysis.

**Constraints on recombination probabilities:** The Minin model applies a conservative constraint to the sum of recombination probabilities restricting the expected number of breakpoints in individual sequences to 0.693. However, application of this constraint makes the assumption that all sequences used in the dataset are of nearly the

same length. For full genome analyses of HIV, much valuable information can be gained from partial genome sequences added to make up for insufficient full length data. Also, as the incidence of recombinants in the global epidemic increases, producing more and more complex recombinants, the assumption of less than 1 recombination breakpoints may be too conservative. In our model, we remove the use of the constraint.

**Ascertainment Bias:** The choice of only recombinants in the dataset can lead to erroneous interpretation of recombination hotspots due to sampling bias. Our model corrects for this sampling bias by weighting probabilities used by individual sequences by their length. The following chapter provides more details and some results.

# CHAPTER 6. DETECTING ASSOCIATION OF RECOMBINATION HOTSPOTS WITH GENOMIC FEATURES IN HIV-1

A paper to be submitted to *Genetics*

Misha L. Rajaram, Drena Dobbs, Susan Carpenter, Vladimir N. Minin and Karin S. Dorman

## Abstract

We present a Bayesian framework for inferring association of genomic features with the spatial distribution of recombination probability from multiple putative recombinants. Recombination confers on retroviruses, the ability to generate chimeric molecules that may boost the replicative fitness of the virus. Many studies have shown that the spatial preferences of recombination breakpoints across HIV recombinants is not uniform, leading to the presence of recombination hotspots. However, much ambiguity exists around the molecular mechanism of recombination and the genomic features that may be associated with recombination hotspots. We propose a hierarchical model that allows for simultaneous inference of locations of recombination breakpoints in multiple sequences, spatial variation of recombination rate in the genome and its association with specified genomic covariates. At the lower level of the hierarchy, a phylogenetic recombination detection model is applied to individual sequences to infer the presence and

location of breakpoints via changepoint processes. At the upper level, recombination probabilities are expressed as a linear function of covariates and spatially varying error terms. We place a spatially smoothing Gaussian Markov Random Field (GMRF) prior on the error term to combine information from the individual sequences. We applied the model to a dataset of 527 simple recombinant sequences covering the HIV-1 genome. We report a putative hotspot in the RT region of the *pol* gene and another in the *nef* gene. We also found positively significant associations of recombination rates with propensity to form secondary structures lending support to the hypothesis that pause sites trigger recombination.

## 6.1 Introduction

The Human Immunodeficiency Virus (HIV) packages its genetic material in two positive sense RNA strands. The lack of an exonuclease proof reading mechanism, high mismatch error rate in the transcription machinery and recombination lead to a high genetic diversity of the virus that help it escape host immune mechanisms. Recombination occurs upon template switching by the reverse transcription machinery. When the co-packaged RNA genomes belong to diverse genomic variants, recombination results in new genetic variants called inter-subtype or inter-specific recombinants. In HIV-1 the frequency of template switching can be anywhere between 7 to 30 times per genome (Levy et al., 2004). Recombinants observed at least three times are characterized as Circulating Recombinant Forms (CRFs); there were 43 known types as of 2009 (HIV-Database, 2010). All other recombinants are called unique recombinant forms (URFs). CRFs have triggered many local epidemics (Frange et al., 2008), and account for an estimated 18% of the global epidemic (Buonaguro et al., 2007) and may gradually outcompete the pure subtypes in the global infecting pool (Tovanabutra et al., 2004; Njai et al., 2006; Wang et al., 2007). Subtypes and CRFs differ in virulence (Baeten et al.,

2006), drug resistance (Spira et al., 2003; Madani et al., 2004) and sensitivity to detection assays (Swanson et al., 2005; Colson et al., 2007).

Increasing experimental evidence suggests that strand transfer events do not occur uniformly along the genome. There have been several hints of a hotspot in the 5' portion of the *pol* gene, both in experiment (Jetzt et al., 2000) and among *in vivo* recombinants sampled from patients (Magiorkinis et al., 2003; Thomson et al., 2004; Galli. et al., 2008). Another well studied hotspot is the conserved C2 region of *env* (Moumen et al., 2001; Quinones-Mateu et al., 2002) although all conserved regions in *env* have relatively high recombination rates (Baird et al., 2006a) and even variable regions are believed to become hotspots with the right donor template (Baird et al., 2006b). Many inter-subtype recombinants observed around the world display a recombinant pattern where the variable loop V3, between C2 and C3 (Renjifo et al., 1999), or the complete *gp120* portion of *env* is swapped with another subtype (Takebe et al., 2003). Other regions implicated as possible hotspots are 5' *gag* (Dykes et al., 2004), the *gag-pol* boundary (Magiorkinis et al., 2003), the *pol-vif* boundary (Derebail et al., 2003), through *vif* into the 5' *env* (Magiorkinis et al., 2003), a GC-rich region near the *tat-rev* splice site (Douglas et al., 1996), and near *nef* (Magiorkinis et al., 2003). Indeed, very few genomic regions are consistently "cold" but few studies have examined the entire genome at once and experimental protocols and reagents vary widely. Additionally, little consensus exists on what causes the replication machinery to fall off a template and on to another. Finding association of recombination hotspots with genomic features may provide important clues to dissecting the mechanism.

Most methods for detection of topology change points use phylogenetic inference (Hein, 1990). These methods exploit the fact that if recombination has occurred in a set of aligned sequences then their phylogenetic relationships should differ on either side of a breakpoint. Most popular approaches, therefore, use a sliding window technique to look for support for alternative topologies (Grassly and Holmes, 1997; McGuire et al., 1997;

Husmeier and McGuire, 2003). This method, however, suffers from a multiple testing problem and low resolution for detecting breakpoints (Suchard et al., 2002). More recent advances have been in the area of a Bayesian Hidden Markov Model (HMM), where the underlying tree topologies are considered the hidden states while the actual observed alignment is considered as the observed state. This method is more accurate than the sliding window methods (Husmeier and McGuire, 2003). However, it is computationally intensive and currently can only include up to four sequences in the alignment. Additionally, it assumes that all regions of the alignment are under similar selection pressures which may be problematic (Dorman et al., 2002; Husmeier and McGuire, 2002).

Change-point models have been used to successfully model the spatial phylogenetic variation along an alignment. A single multiple change point model (Suchard et al., 2003) was extended to a dual multiple changepoint (DMCP) Model (Minin et al., 2005) that successfully de-convolutes changepoints arising from recombination breakpoints from other types of changepoints along an alignment. The DMCP achieves this by modeling changes in nucleotide substitution pressures and tree topology as two independent changepoint processes thus providing more accuracy to the recombination detection problem by decoupling the effect of nucleotide substitution from real tree topology change.

Minin et al. (2007) introduced a Bayesian hierarchical model to combine breakpoint information from many individual sequences in order to infer recombination hotspots at a genome level. This model achieves such inference by assuming a genome level recombination probability that informs on the placement of breakpoints in individual sequences. A spatially smoothing intrinsic Gaussian Markov Random Field (GMRF) prior is placed on the vector of recombination probabilities. This prior defines the neighborhood of each site, sites to its immediate left and right, enabling spatial smoothing by allowing the sharing of information between neighbors. While pioneering in its application of the GMRF prior to the inference of recombination hotspots in HIV, assumptions made in

the above model preclude the use of a large number of sequences as also sequences of disparate lengths.

In the present paper we describe a Bayesian hierarchical model that simultaneously infers recombination hotspots and their association with genomic features. The lower lever of the model infers topology change points in individual sequences using the DMCP model from (Minin et al., 2005). The genome level counts for site-wise topology change-points are then used to capture the recombination probability via a binomial likelihood function. The upper level of the hierarchy makes inference on the recombination probabilities as well as the regression coefficients measuring their association with input covariates. A one-dimensional GMRF prior (Minin et al., 2007) is placed on the recombination probabilities to enable smoothing of recombination probabilities. Further, in order to take full advantage of all the publicly available HIV sequence data, we extended the model so it is not only able to handle a large number of sequences but also sequences of variable length. Our model also corrects, automatically, for any sampling bias that the input dataset may introduce.

Section 2 provides an overview of the model as well as methodology adopted in curation of the test dataset. Section 3 presents results of simulation runs with artificial recombinant sequences of varying lengths and inclusion of different types of covariates. The latter part of this section presents results from application of the model to a dataset containing known HIV-1 recombinants and their association with genomic covariates. We included GC content and sequence similarity as well as features associated with RNA secondary structure formation. We find that the latter have a significantly positive association with recombination probabilities, i.e. those that have been hypothesized to promote formation of secondary structures also seem to be associated with the presence of recombination hotspots.

## 6.2 Methods

### 6.2.1 Mapping Sequences to HXB2

Minin et al. (2007) map individual sequences to a population level vector by using a multiple sequence alignment containing all sequences in the dataset along with their putative parental sequences. However, this requirement becomes increasingly prohibitive with larger datasets. We solve this problem by mapping individual sequences back to HXB2 positions disallowing any gaps in the HXB2 sequence. While we lose information about insertions relative to HXB2, the gaps are not long and this strategy does away with the need for a multiple sequence alignment thus allowing for the inclusion of many sequences in our dataset.

Let $\boldsymbol{Y}$ denote the data of $K$ individual alignments, $(Y_1, Y_2 \dots Y_K)$, each corresponding to a multiple sequence alignment containing a putative recombinant and its parental sequences. Note that gaps in the recombinant are removed from these alignments since recombination breakpoints can not be placed at these gap sites. Individual sequences may then be mapped to their corresponding HXB2-relative positions using a mapping function:

$$f_k : 1, 2....L_k \rightarrow 1, 2...S$$

where $L_k$ is the length of alignment $Y_k$, $S$ is the total length of the region of the HIV genome covered by the dataset of $K$ sequences and $f_k(i)$ is the HXB2 site corresponding to site $i$ in this alignment.

### 6.2.2 The Dual Multiple Changepoint Model (DMCP)

The lower level of the hierarchical model, infers recombination breakpoints in individual alignments $Y_k$ via the dual multiple changepoint (DMCP) model (Minin et al., 2005). Columns in the alignment $Y_k = (Y_k^{(1)}, Y_k^{(2)}, \dots, Y_k^{(l)}, \dots, Y_k^{(L_k)})$ are assumed to evolve independently as a continuous time Markov chain with transition/transversion

ratio $\kappa_k^{(l)}$, following Hasegawa et al. (1985). The stationary distribution parameters $\pi_{kN}, N \in \{A, C, G, T\}$ are fixed to the observed proportions in the input alignment. We complete the specification of the phylogenetic model by specifying a bifurcating tree topology $\tau_k^{(l)}$ that models the evolutionary relationship of the nucleotides in column $Y_k^{(l)}$. An exponential prior with mean $\mu_k^{(l)}$ is placed on the branch lengths of the tree, to reduce the number of free parameters (Suchard et al., 2003; Minin et al., 2005). This specification leads to a site-wise likelihood,

$$\mathcal{L}_k^{(l)} = P(Y_k^{(l)} | \tau_k^{(l)}, \mu_k^{(l)}, \kappa_k^{(l)}) \tag{6.1}$$

A change point may occur when there is a change in the tree topology $\boldsymbol{\tau_k}$, a change in the evolutionary parameters $(\boldsymbol{\mu_k}, \boldsymbol{\kappa_k})$, or both. Let $1 = \theta_0 < \theta_1 < \cdots < \theta_{M_k+1} = L_k + 1$ represent the locations of $M_k$ unknown and distinct topology change points. All positions in the alignment between adjacent topology change points, i.e. $l \in [\theta_{m-1}, \theta_m), 1 \leq m \leq M_k + 1$ share a tree topology $\tau_k^{(m)}$, and adjacent fragments have distinct topologies. Let $1 = \rho_0 < \rho_1 < \ldots < \rho_{J_k+1} = L_k + 1$ be the locations of $J_k$ distinct evolutionary changepoints that mark changes in evolutionary parameters. As before, all sites of the alignment between two evolutionary changepoints share the same evolutionary parameters.

The DMCP model is, therefore, completely defined by parameters

$$\Phi_k = \{M_k, J_k, \boldsymbol{\theta_k}, \boldsymbol{\tau_k}, \boldsymbol{\rho_k}, \boldsymbol{\mu_k}, \boldsymbol{\kappa_k}\}.$$

Our interest is in the number and location of topology change points $(M_k, \boldsymbol{\theta_k})$. Hence, we bundle the rest of the parameters as nuisance parameters in vector $\boldsymbol{\psi_k}$. Priors on nuisance parameters were set as described in Minin et al. (2007).

### 6.2.3 Prior on Location of Topology Changepoints

The upper level of the hierarchy combines information from the $K$ recombinants analyzed in the lower level of the hierarchy. We reconfigure the topology changepoints

$\boldsymbol{\theta}_k$ of recombinant $k$ into a vector of indicator variables $\boldsymbol{B}_k = (B_{k,1}, B_{k,2}, \ldots, B_{k,L_k})$ where $B_{k,l} = 1$ if $l \in \boldsymbol{\theta_k}$, 0 otherwise. Let $p_s \in \boldsymbol{p} = (p_1, p_2 \ldots p_S)$ be the population-level probability that site $s$ is a topology change point. Then, the probability of the current configuration of topology change points in alignment $k$ is

$$Pr(B_{k,l} = r|\boldsymbol{p}) = (p_{f_k^{(l)}})^r (1 - p_{f_k^{(l)}})^{1-r}, \text{ where } r = \{0, 1\}. \tag{6.2}$$

Conditional on the recombination probabilities $p_s$, we assume that topology change points are independent, so the joint likelihood of $\boldsymbol{B}$ is

$$Pr(\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots \boldsymbol{B}_K|\boldsymbol{p}) \propto \prod_{k=1}^{K} \prod_{l=1}^{L_k} (p_{f_k^{(l)}})^{B_{k,l}} (1 - p_{f_k^{(l)}})^{1-B_{k,l}}. \tag{6.3}$$

Denoting all recombinants that did not have a gap character at site $s$ as $C_s = \sum_{k=1}^{K} \{s \in \text{range}(f_k)\}$ and the total number of times site $s$ was inferred to be breakpoint as $R_s = \sum_{k=1}^{K} B_{k,s}$, (6.3) simplifies to

$$Pr(\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots \boldsymbol{B}_K|\boldsymbol{p}) \propto \prod_{k=1}^{K} p_s^{R_s} (1 - p_s)^{C_s - R_s} \tag{6.4}$$

In order to infer recombination probabilities from the likelihood in (6.4), it is imperative to define an informative prior structure on them. An efficient prior that is also biologically relevant is the Gaussian Markov Random Field (GMRF) prior (Minin et al., 2007). We first obtain the logit transformation of the recombination probabilities,

$$\nu_s = \log\left(\frac{p_s}{1 - p_s}\right) \tag{6.5}$$

and apply the GMRF prior to the recombination log-odds.

$$\boldsymbol{\nu}|\omega \sim \omega^{(S-1)/2} \exp\left(-\frac{\omega}{2} \sum_{s=1}^{S-1} (\nu_s - \nu_{s+1})^2\right) \tag{6.6}$$

This prior penalizes large differences in recombination probabilities at adjacent sites. We refer interested readers to Minin et al. (2007) for more details on the structure of the precision matrix as well as adjustments made to ensure that the resultant posterior is proper.

### 6.2.4 Modeling Association of Genomic Features

In order to extend the current model to include covariates of interest, we express the recombination log-odds $\boldsymbol{\nu}$ as

$$\nu_s = \eta_s + X_s\beta \tag{6.7}$$

where $\boldsymbol{X}$ denotes the model matrix, $\boldsymbol{\beta}$, the set of regression coefficients and $\boldsymbol{\eta}$ is the spatially varying component of $\boldsymbol{\nu}$. Note that $\boldsymbol{\nu}$ is still a GMRF, now with mean $\boldsymbol{X\beta}$ and dimension $S + n$ where $n$ is the dimension of $\boldsymbol{\beta}$ (Rue and Held, 2005).

We complete the Bayesian formulation of the model by specifying a non-informative uniform prior on $\boldsymbol{\beta}$ and by fixing $\omega$ appropriately. In fact, $\omega$ can be viewed as a smoothing parameter that can be adjusted to match the availability of data. Note that $\omega$ may be inferred as well, in which case we place a $\Gamma(\omega_a, \omega_b)$ prior on it where $\omega_b$ is fixed at 0.02 and $\omega_a$ at $S - 1$. For more details, we refer interested readers to Minin et al. (2007) and Rue and Held (2005).

### 6.2.5 Inference via MCMC Simulation

The joint posterior distribution of all model parameters is

$$\begin{aligned} &Pr(\Phi_1, \Phi_2 \dots \Phi_K, \boldsymbol{\nu}, \omega \mid Y_1, Y_2 \dots Y_K) \\ &\propto \prod_{k=1}^{K} Pr(Y_k \mid \Phi_k) Pr(\psi_k) Pr(R_k \mid \boldsymbol{\nu}) \times Pr(\boldsymbol{\eta} \mid \omega) Pr(\boldsymbol{\beta}) Pr(\omega). \end{aligned} \tag{6.8}$$

In order to sample from (6.9), we employed MCMC simulation via the block update scheme proposed by Rue and Held (2005) for sampling from distributions involving GMRF hyperpriors. Model parameters were updated in three blocks. In the first block, we updated parameters of individual alignments by sampling from their full conditionals.

$$Pr(\Phi_1, \Phi_2 \dots \Phi_K, \mid \boldsymbol{\nu}, \omega, Y_1, Y_2 \dots Y_K) \propto \prod_{k=1}^{K} Pr(Y_k \mid \Phi_k) Pr(\psi_k) Pr(R_k \mid \boldsymbol{\nu}) \tag{6.9}$$

Note that the $\Phi_k$ are conditionally independent, given $\boldsymbol{\nu}$ and $\omega$ and simulation simply involves cycling through all individual alignments. Minin et al. (2005) used an rjMCMC

sampler to update parameters of individual alignments. In the current model, we retain their sampling scheme with appropriate adjustments for the prior on locations of topology changepoints.

At the upper level of the hierarchy, the full conditional is

$$Pr(\boldsymbol{\eta}, \boldsymbol{\beta}, \omega \mid Y_1, Y_2 \ldots Y_K, \Phi_1, \Phi_2 \ldots \Phi_K) \propto Pr(\boldsymbol{R}, \boldsymbol{C} \mid \nu) Pr(\boldsymbol{\eta} \mid \omega) Pr(\boldsymbol{\beta}) Pr(\omega) \quad (6.10)$$

From (6.4) and (6.5), we may express $Pr(\boldsymbol{R}, \boldsymbol{C} \mid \nu)$ as

$$Pr(\boldsymbol{R}, \boldsymbol{C} \mid \nu) \propto \prod_{s=1}^{S} \left( \frac{e^{\nu_s}}{1 + e^{\nu_s}} \right)^{R_s} \left( \frac{1}{1 + e^{\nu_s}} \right)^{C_s - R_s} \quad (6.11)$$

The second block consists of the spatial parameters $\boldsymbol{\eta}$ and $\omega$ that are updated jointly following a strategy proposed by Rue and Held (2005). In the current model, however, we fix $\omega$, thereby reducing the parameter space of this block only to $\boldsymbol{\eta}$. The third block includes the $\boldsymbol{\beta}$ coefficients that are updated independent of the $\boldsymbol{\nu}$ or $\omega$ updates.

From (6.10), the posterior marginal distribution of $\boldsymbol{\nu}$ can be denoted as

$$Pr(\boldsymbol{\nu} \mid \boldsymbol{R}, \boldsymbol{C}, \omega) \propto \exp(-\frac{1}{2} \boldsymbol{\nu}^T \boldsymbol{Q} \boldsymbol{\nu} + \sum_{s=1}^{S} R_s \nu_s + C_s \ln(1 + e_s^{\nu})). \quad (6.12)$$

We use (6.12) to obtain full conditional for $\boldsymbol{\eta}$ as

$$Pr(\boldsymbol{\eta} \mid \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{C}, \omega) \propto \exp(-\frac{1}{2} \boldsymbol{\eta}^T \boldsymbol{Q} \boldsymbol{\eta} + f(\boldsymbol{\eta}; \boldsymbol{R}, \boldsymbol{C})) \quad (6.13)$$

where $f(\boldsymbol{\eta}; \boldsymbol{R}, \boldsymbol{C})$ is the part of the likelihood pertinent to the full conditional of $\boldsymbol{\eta}$. Similarly, the full conditional for $\boldsymbol{\beta}$ can be denoted by $f(\boldsymbol{\beta}; \boldsymbol{R}, \boldsymbol{C})$. Note that due to the non-normal nature of the likelihood functions, these full conditionals are analytically intractable. To sample from them, we use the Metropolis-Hastings algorithm with proposal distributions described below.

Applying the second order Taylor series expansion to $f(\boldsymbol{\eta}; \boldsymbol{R}, \boldsymbol{C})$, we obtain a GMRF approximation to (6.13) that may be used as a proposal distribution. Note that the Taylor series expansion can be performed around the mode of (6.13) or the current state

$\boldsymbol{\eta}$. The proposal $\eta^*$ thus generated is accepted/rejected using the standard acceptance ratio procedure of the Metropolis-Hastings algorithm (Hastings, 1970). In cases where $\omega$ is also estimated, the proposals $\eta^*$ and $\omega^*$ are jointly accepted/rejected. Arbitrary normal proposal distributions were employed to update each of the members of $\boldsymbol{\beta}$ using series of Metropolis-within-Gibbs steps.

### 6.2.6 Dataset

In order to study the association of recombination probability with sequence covariates, we curated a dataset of 527 recombinants. To gain insight into features associated with the mechanism itself, we curated the dataset in such a way that there were no replicate representatives of a single recombination event. Starting with $2,360$ simple recombinants, i.e those with only two parental genotypes, we grouped these into 544 clusters based on their *profile structure*. The *profile structure* of a recombinant refers to the $5'$ to $3'$ listing of recombination breakpoints and the parental genotypes for every segment created thereof. Rajaram et al. (2007) provides a detailed description of the clustering algorithm. Because the HIV reference dataset used for parental genotype does not have $5'$ LTR sequences, we eliminated sequences that covered only this region and truncated others that extend into the LTR region leading to a dataset of 527 recombinants. Lengths of the recombinants range between 400 and 8587 nucleotides covering 8924 nucleotides corresponding to the entire genome from the $3'$ end of the $5'$ LTR.

### 6.2.7 Correcting Ascertainment Bias

Ascertainment or sampling bias may result in erroneous inference of hotspots. In the current context, this bias may arise when datasets are curated to contain only recombinants and a portion of these only cover a short region of the genome. Repeated inference of breakpoints from these recombinants may result in a posterior probability profile that appears to have a hotspot in high coverage regions. To correct for the

sampling bias, we use bias coefficients specific to each sequence. For recombinant $k$ covering region $(l_k, L_k)$ of the genome, we compute the bias corrected recombination probabilities $p^*_{f_k(s)}$ as

$$p^*_{f_k(s)} = A_k p_{f_k(s)} \tag{6.14}$$

where $A_k$, the recombinant-specific bias coefficient, is computed as

$$A_k = \frac{1}{1 - \prod_{s \in (l_k, L_k)} (1 - p_{f_k(s)})} \tag{6.15}$$

## 6.3 Results

### 6.3.1 Simulation Study

To demonstrate the working of our model in detecting recombination hotspots and their association with covariates, we designed a few simulation cases. We start with a set of $K$ simulated recombinants covering a region of 8121 sites corresponding to the *gag*, *pol* and *env* genes of the HIV genome created using 80 non-recombinant full length sequences each of the B and C genotypes. To test that the model handles sequences of variable length, 25% randomly chosen sequences from this set were truncated so as to cover only the *gag* gene and 25% others cover only the *env* gene. We set the "true" recombination probability for these recombinants such that the region between sites $3500 - 4500$ had maximum probability of having a breakpoint, thus creating a hotspot in the dataset. When a covariate was included, it was part of the "true" probability profile used to create the simulated recombinants. Note that these probability profiles are scaled so that the expected number of changepoints in a full length sequence is 0.693 so as to mimic simple recombinants (Minin et al., 2005). As a consequence, some sequences were not recombinant and no ascertainment bias correction is needed. We report 95% Bayesian credible intervals for the coefficients of regression of the covariates. Coefficients whose credible interval includes 0 are said to have no association with the

recombination probability.

We generated a dataset of 100 simulated recombinants as described above. The probability profile used placed a hotspot in the $3500 - 4500$ region and had no included covariates. The top left plot of Fig. 6.1 shows the simulated probabilities used to create the recombinants. Solid dots on the plot mark the locations of breakpoints in the dataset. As a control case, an unrelated covariate term was provided to the model during inference. The top right plot in Fig. 6.1 shows the posterior inferred recombination probabilities. It also indicates the 95% BCI for the coefficient of the unrelated coefficient, with the middle number indicating the mean and the left and right flanking numbers indicating the lower and upper bounds of the credible interval respectively. The use of sequences of variable length does not affect the accurate inference of the hotspot. Note also that BCI of the coefficient includes 0, i.e. no evidence of association with the covariate, as is truly the case.

Next we tested a case where we included a normal random covariates in the recombination probabilities used to create the test dataset. $\beta$ was fixed at 1.0. The "true" recombination probabilities still placed a hotspot in the region as above, however, the profile of this density was not smooth as before. Fig. 6.1(c) shows the probability profile with added covariate and the locations of breakpoints in the dataset of 100 simulated sequences created from it. Fig. 6.1(d) shows the posterior inference of recombination probabilities. The directionality of the $\beta$ term was inferred correctly although its inferred strength is diminished.

Finally, to test the scenario where the covariate itself may present with a continuously varying structure, we used a probability profile as shown in Fig. 6.1(e). The first "hotspot" region is created by the covariate, with $\beta = 1.0$ while the second hotspot in the $3500 - 4500$ region is congruent with the ones in prior datasets and corresponds to the spatial error term. Fig. 6.1(f) shows the inferred recombination probabilities. The 95% BCI for $\beta$ was significant and positively associated. However the mean is much

lower than the true value of $\beta$. We tested two more datasets, one with 250 sequences and a third with 500 sequences and found that the mean of the inferred coefficient is closer to the true value with more data.

To demonstrate the use of the ascertainment bias correction coefficients, we created a simulated dataset of 100 recombinants. As before, 25 each of these covered the *gag* and *env* regions and 50 were full length sequences. Here, we set the expected number of breakpoints for each sequence, regardless of length, to 1 and these were uniformly distributed along the genome resulting in no hotspots in the dataset. Fig. 6.3(a) shows the posterior inference on this dataset without bias correction. Solid dots mark locations of recombination breakpoints in the dataset. As predicted earlier, the recombination profile shows a distinct pattern similar to the coverage pattern. Fig. 6.3(b) shows bias-corrected posterior inference accurately reflecting the uniform distribution of breakpoints in the dataset.

### 6.3.2 Analysis of Real Data

The model was used to analyze the dataset curated as described in section 6.2.6. Note that since all the sequences were chosen to be recombinants, we applied the ascertainment bias correction for this analysis. Figure 6.4 shows the posterior recombination probability inferred from this dataset. The posterior mean probabilities are plotted (line) along with the 95% credible interval (shading). Lack of parental sequence information in the 5' LTR region caused us to exclude this region. The x-axis is numbered relative to HXB2 positions. A hotspot occurs in the $5'$ end of the *pol* gene coding for the reverse transcriptase enzyme. Another hotspot occurs in the *gag* gene. The $3'$ end of the *pol* gene coding for protease is particularly cold. The *env* gene tends to have moderately high recombination rate throughout. A peak is observed in the region overlapped by *nef* and the $3'$ LTR.

The next few sections present results of associations with various covariates. Table 6.1

provides a summary of the mean and 95% BCI for the association coefficients. It should be noted that all covariates were normalized before inference and as such the magnitude of the coefficients obtained are directly comparable to each other.

**Association with known recombination hotspots:** Many previous studies (Moumen et al., 2001; Dykes et al., 2004; Balakrishnan et al., 2003; Zhuang et al., 2002; Baird et al., 2006b) have, through *in vitro* systems or single cell infection assays, identified recombination hotspots along the HIV-1 genome. To test how these translate to *in vivo* sequences where they face selection pressure, we included as covariate an indicator variable that had a value 1 for any nucleotide that was part of the hotspots identified by Moumen et al. (2001); Zhuang et al. (2002); Dykes et al. (2004) and Baird et al. (2006b), and 0 otherwise. We found no significant association with these hotspots.

**GC content and Sequence Similarity:** Rajaram et al. (2007) examined the correlation between GC content and sequence similarity with posterior recombination probabilities. Here we included these as covariates in the model. GC content was summarized using sliding windows of sizes 20, 50 and 100 bp. The GC content at site $s$ is the proportion of G and C nucleotides within a window of size $n$ centered at site $s$.

Sequence similarity was summarized using Shannon's entropy. For site $s$ the entropy is computed as

$$H_s = -\sum_{i\in(A,C,G,T)} \rho_{(s,i)} \log(\rho_{(s,i)}) \tag{6.16}$$

where $\rho_{(s,i)}$ corresponds to the proportion of appearance of nucleotide $i$ at site $s$ in the entire dataset. Further Magiorkinis et al. (2003) reported association with upstream sequence similarity and frequency of recombination. To summarize this, we also considered entropy covariates in sliding windows of size 10, 20 and 50 bp. In this case the entropy at site $s$ was computed as the average positional entropy of the nucleotides occurring in the window starting at that site.

The association of GC content is positive and increases with increasing window size. Site-wise entropy has a negative association with recombination probability while window-wise entropy has a positive association. The strength of window-wise entropy associations also increases with larger windows.

**Deletions:** The jumping reverse transcription complex, could introduce insertions and deletions in the resultant recombinant. While our use of HXB2 mapping of individual sequences prevents the analysis of insertions, we may use deletion counts to test their association with recombination probabilities. The deletion count at site $s$ is the length of the gap region that follows it. The mean of the coefficient was inferred very close to 0 with its 95% BCI including 0.

**Thermodynamic stability of RNA/DNA hybrid:** Sugimoto et al. (1995) provided parameters for the nearest neighbor computations to predict the thermodynamic stability of a RNA/DNA hybrid. The covariate at site $s$ was the total free energy, $\Delta G_{37\,\mathrm{deg}}$, of the 9-mer centered at site $s$ computed using the nearest neighbor method. The association of this covariate with recombination probability was inferred to be significant and negative.

**Covariates associated with RNA Secondary structure:** Next we examined association of some covariates that indicate formation of secondary structures by the genome RNA. Many studies (Galetto et al., 2006; Moumen et al., 2003; Roda et al., 2002b) have indicated association of secondary structure with recombination and hypothesize that a loop may act as a pause site for the reverse transcription machinery, encouraging it to fall off the template genome. Shen et al. (2009) showed that G-rich stretches along the HIV genome are prone to formation of tetraloops and associated with recombination. To include the presence of G-rich regions as a covariate we adopted the following coding system. All nucleotides other than G were coded 0. Guanosine

nucleotides were coded in increasing order, i.e. the first in a stretch of Gs was coded 1, the next 2 and so on until the stretch was interrupted with a non-G nucleotide, resetting the counter to 0. The coefficient for this covariate is positive and significant.

Watts et al. (2009) recently elucidated the secondary structure of the HIV genome using the high-throughput selective 29-hydroxyl acylation analysed by primer extension (SHAPE) (Wilkinson et al., 2006; Deigan et al., 2009) method. SHAPE reactivity measures the propensity of a nucleotide to be acetylated. It, therefore, provides very clear information of whether a nucleotide is involved in a secondary structure or not. High SHAPE reactivities indicate unstructured nucleotides or absence of secondary structures. They also use a method suggested by (Pedersen et al., 2004) to compute pairing probabilities at each nucleotide. This algorithm does not involve chemical or thermodynamic computations and has opposite directionality to the SHAPE reactivities, i.e. high pairing probability indicates higher propensity for the nucleotide to be part of a secondary structure. We provided the SHAPE reactivity and pairing probabilities as covariates and found that the resultant coefficients were significant. SHAPE reactivity was negatively associated with recombination probability while pairing probabilities were positively associated.

### 6.3.3 MCMC Convergence Diagnostics

At the lower level, individual scaled regeneration quantile (SRQ) plots for the total number of inferred breakpoints $M_k$ were examined to assess convergence. For the DMCP model, the time evolution of $M_k$ is a vital indicator of mixing in the reversible jump MCMC sampler. Mykland et al. (1995) suggest that for renewal process such as the MCMC sampler of the DMCP model, plotting its regeneration times can be used to assess convergence. $M_k$ has a discrete state space and we mark the state denoting the posterior median number of breakpoints as the state of interest $m$. For an MCMC run of fixed length, $t_i, i = 1, 2 \ldots n$ are the time steps at which state $m$ was visited where

$n$ is the random total number of visits to state $m$. A plot of $t_i/t_n$ vs. $i/n$ close to the diagonal corresponding to the $y = x$ line indicates convergence. Fig. 6.5 shows the SRQ plots for all 527 individual sequences in our real dataset with all lines lying very close to the diagonal.

We employ the Geweke statistic (Geweke, 1992) to test convergence at the upper level. The first 10% and last 50% samples were used for this test. Geweke suggests that treating the samples as a time series, the Z-statistic for the difference of the means thus computed is asymptotically normal. A Z-test was performed for mean $\nu$ at each position and mean $\beta$ for the two samples. No difference was found significant thus indicating convergence.

## 6.4 Discussion

We presented a hierarchical model that can assess, for a set of putative recombinants, presence of recombination hotspots and their association with covariates. Extending the model presented by Minin et al. (2007), we are now able to deal with large datasets that may also include sequences of varying lengths. As the simulation results indicate, providing more information in terms of a larger number of sequences can greatly help in achieving better resolution of the inference. Further, the model also handles bias arising from sampling a subset of sequences. This bias becomes especially acute when sequences with varying lengths are allowed. In trying to understand the molecular mechanism of recombination, care must be taken to ensure that recombinants chosen do not represent replicates of the same recombination event. Covariate associations found in datasets with replicates while informative on replicative fitness of a recombinant, do not provide insights into the mechanism itself. Abundance of a particular recombinant in the dataset will result in heightened recombination probabilities in areas near its breakpoints leading to spurious hotspot inference. The bias correction included in the current model handles

this by weighting the recombination probabilities at the lower level by the length of the individual sequence.

We used a carefully curated dataset to find associations with sequence features. Analysis of this dataset reveals that considerable spatial variation in recombination probability along the HIV genome. Galli. et al. (2008) hypothesized the presence of a recombination hotspot through a computational study that compared frequency of breakpoint occurrence to frequency of occurrence of hairpin loops in the *pol* gene. The advent of drug therapy and the subsequent drug resistance mutations that have accumulated in regions that are drug targets, such as RT could be a cause for this hotspot (Charpentier et al., 2006; Nora et al., 2007). However, drug therapy itself is not very prevalent in poorer sections of the world (WHO, 2007) and may not fully explain the hotspot. A hotspot in the *gag* gene was reported in an experimental study (Shen et al., 2009). Minin et al. (2007) also reported a putative hotspot in the *gag* gene from the application of the hierarchical GMRF model to a set of A/G recombinants. The peaks in the *env* gene occurred at the edges of the *gp120* coding region. The surface glycoprotein, *gp*120, is the first line of attack for the virus and switching it out with that of a different genotype may confer evolutionary advantage. Indeed, in some A/E recombinants from Thailand, the *gp*120 coding region is from subtype E while the rest of the virus is subtype A1 (Sabino et al., 1994). This is also a highly successful recombinant and one of the major infecting types of the region. Evidence of recombination in the *nef* gene was earlier reported in *in vitro* studies of SIV. More recently, inter-subtype *env-nef* recombinants have been reported in India (Bhanja et al., 2007).

### 6.4.1 Quantifying Associations with Genomic Features

We performed a series of experiments with simulations to test our model for most plausible scenarios. $\beta$ was underestimated when a random covariate is included. This may be attributed to the spatial term accounting for some of the random noise added by

the covariate. Estimates for covariates of this nature are thus bound to be conservative. Further, we found that adding more sequences to the dataset results in more lucid posterior inferences. An advantage our model has over previous models is its ability to test a large number of sequences simultaneously.

Much debate exists on the molecular mechanism of recombination, specifically the triggers that prompt the replication machinery to fall off a donor template, the "signals" that help it anneal to a receptor template and the many interactions involved in between. Galetto and Negroni (2005) provide an extensive review of evidence to support the copy choice model for recombination (Hu and Temin, 1990). Magiorkinis et al. (2003), through a combination of *in vitro* and computational studies, provide evidence that they interpret as support for strand displacement-assimilation model. Many other *in vitro* studies have found association of recombination with RNA secondary structures (Galetto and Negroni, 2005; Moumen et al., 2003; Roda et al., 2002b), GC rich regions (Klarmann et al., 1993) etcetera. However, how these forces act *in vivo*, if at all, has been difficult to measure. This model provides a paradigm for simultaneous inference of recombination hotspots and coefficients of association.

Recent studies have demonstrated the presence of recombination hotspot along the HIV-1 genome (Zhuang et al., 2002; Balakrishnan et al., 2003; Dykes et al., 2004; Galetto et al., 2006; Moumen et al., 2001). However, these studies were performed using reconstituted *in vitro* systems or using single infection assays *in vivo*. It must be noted that sequence data available from patients constitutes viruses that have undergone selection pressure and hence these may not exhibit the hotspots found in the studies above. Inclusion of a covariate that measures association of recombination hotspots identified by (Moumen et al., 2001; Zhuang et al., 2002; Dykes et al., 2004; Baird et al., 2006b) results in a coefficient with no effect. This suggests that selection plays a role in the evolution of the virus within infected individuals.

GC content is one of the simplest features summarizing the genome. In the current

study we find a positive and significant association of recombination probability with GC content and it is found to increase with larger windows. However, this association may possibly reflect the varying GC content across the genome itself. Klarmann et al. (1993) reported that the transcription machinery pauses at GC rich stretches and this correlation may be indicative of the association of pauses and recombination.

Entropy is also found to have significant association with recombination probability. Per-site entropy had a negative association, supporting the theory that some amount of homology is necessary at recombinogenic regions. However, we find that window-wise entropy had the opposite effect. While it is possible that variability may be associated with high recombination rates, a more plausible explanation is that these reflect the mechanism of inference of breakpoints rather than a true association. The lower level DMCP model, by virtue of being a phylogenetic recombination detection model requires some amount of variation to be able to place recombination breakpoints in a region.

The strength of the RNA/DNA hybrid may contribute towards keeping the transcription machinery on the donor template, thus preventing recombination. The significant negative association to this covariate supports this hypothesis. This effect is orthogonal to the effect of secondary structures. Secondary structures have been hypothesized to cause the transcription machinery to pause or sometimes prevent it from moving further (Dykes et al., 2004; Lanciault and Champoux, 2006). Such pause sites may provide an impetus to recombination events (Roda et al., 2002a, 2003; Zhuang et al., 2002). We summarized the propensity for secondary structure formation using three covariates. Shen et al. (2009) hypothesized that the presence of guanosine rich runs may promote the formation of secondary tetra-guanosine loops. A significant positive association of this covariate with recombination probability was observed. Watts et al. (2009) provide details of SHAPE reactivities and pairing probabilities of the HIV-1 genome. Regions with high SHAPE reactivities are considered to be unstructured loops while those with medium- to low-random reactivities may be members of a loop. Similarly, pairing proba-

bilities measure the propensity of a nucleotide to be part of a secondary structure. Lower pairing probabilities indicate unstructured regions. Significant negative and positive associations were observed with SHAPE reactivity and pairing probabilities respectively.

The factors promoting recombination *in vivo* are largely unknown (Negroni and Buc, 2001). This model provides a unified framework to infer these associations. The covariates in turn augment the recombination hotspot prediction by infusing more data into the model. Overall, our model is superior to previous efforts that use phylogenetic recombination detection to infer spatial variation of recombination probability and its association with sequence features (Magiorkinis et al., 2003; Zhang et al., 2005). Using the hierarchical setup it is able to integrate over all recombination breakpoints inferred at the lower level.

The model may also be used to comment on epidemiological impact of recombination. Using datasets representative of specific time periods in the evolution of the virus, comparative studies can be carried out to examine if and how the spatial variation of recombination rates differs between the two time periods. Covariates such as presence of drug resistance mutations may be included to test if they share a significant association with the change in recombination probability profile. Finally, the model may be extended to include covariates at the level of the individual sequence. This extension will enable testing of some very interesting temporal and geographical hypotheses such as, how the spatial recombination profile has changed over time, how recombinants from one geographic region compare with those from another in propensity to recombine and placement of hotspots, how genotypes differ in these aspects and so on.

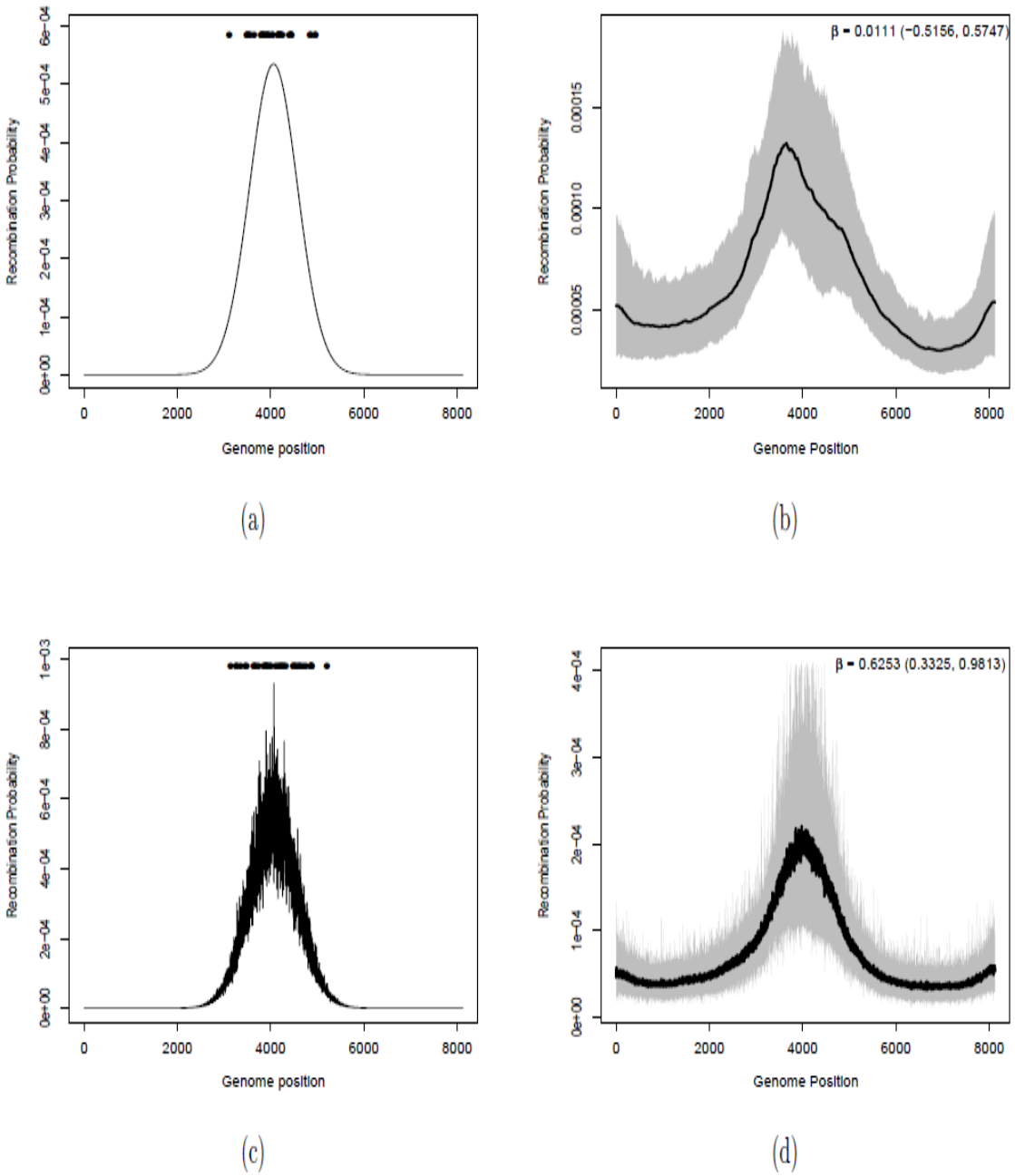



Figure 6.1 (a) Recombination probability profile with no included covariate (b) Posterior inference for dataset generated in (a) analyzed with an unrelated covariate. (c) Recombination probability profile with including a randomly varying covariate (d) Posterior inference for dataset generated using profile in (c).

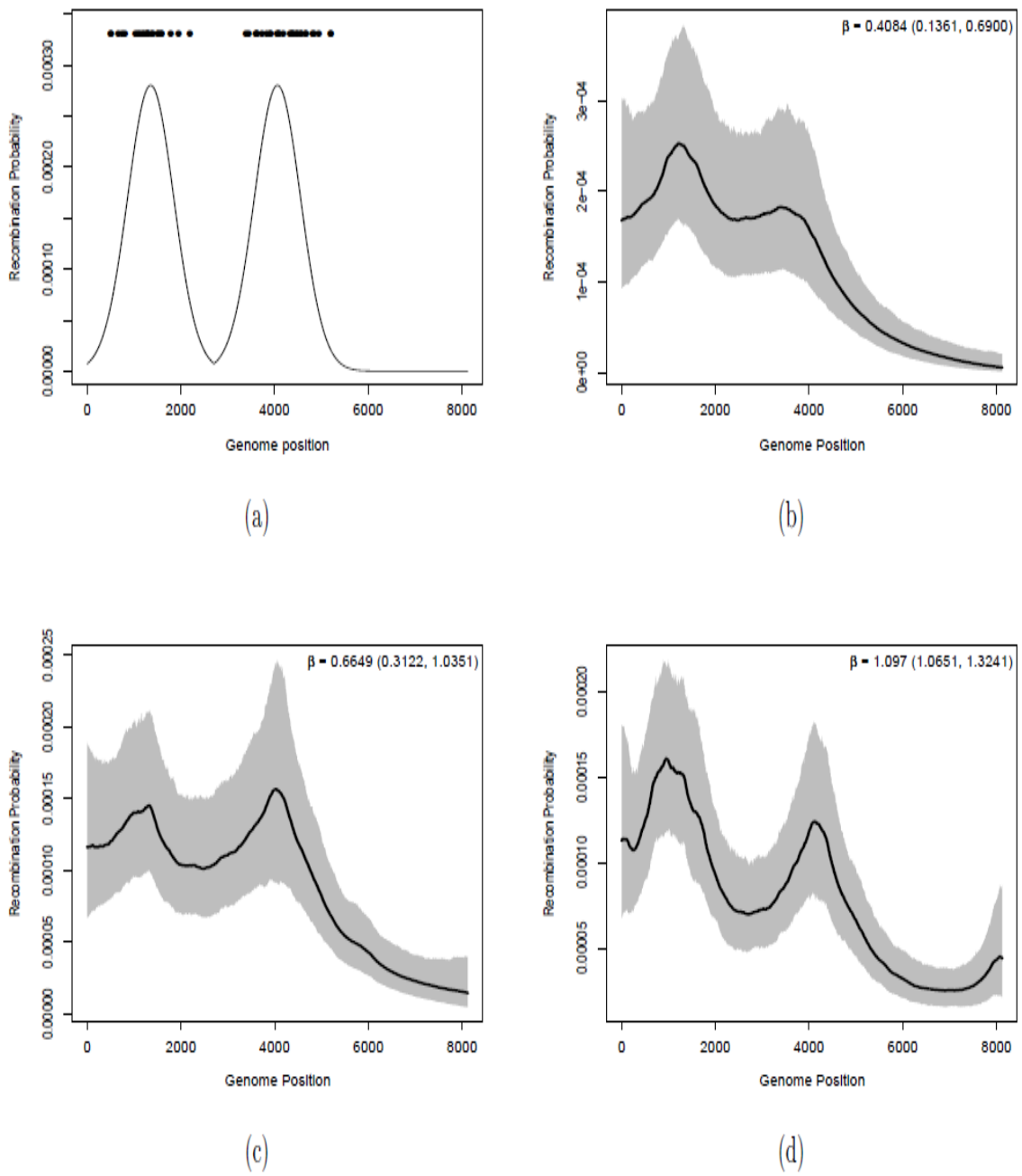



Figure 6.2 (a) Recombination probability profile with spatially structured covariate (b-d) Posterior inference for dataset with 100,250 and 500 simulated recombinants respectively, generated using profile in (a).

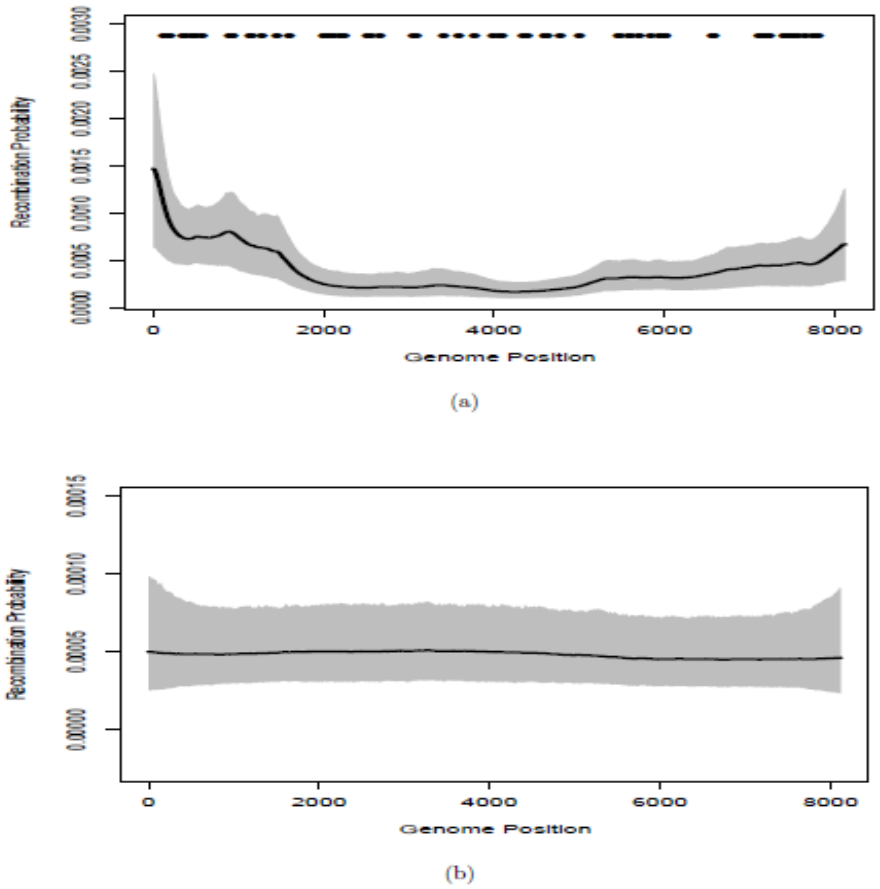



Figure 6.3 Results from simulation dataset to test correction of ascertainment bias (a) Posterior inference before bias correction. Solid dots indicate locations of breakpoints in the dataset. (b) Posterior inference after applying bias correction

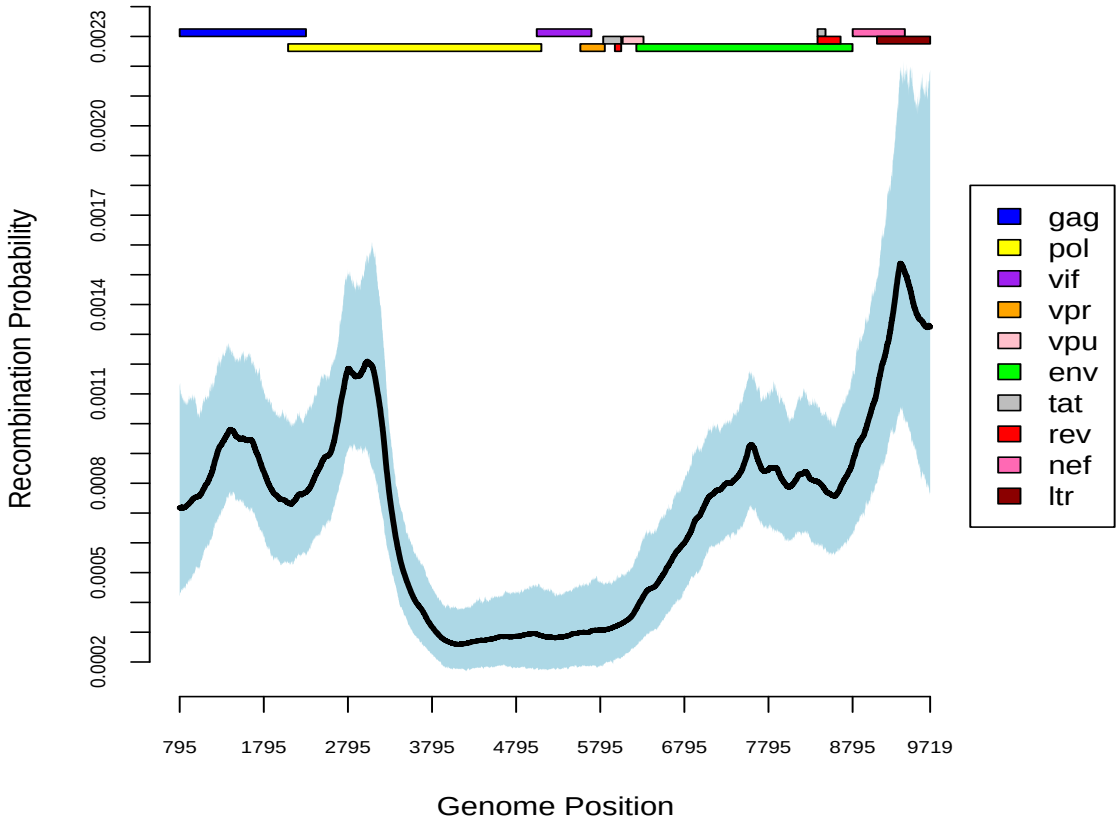



Figure 6.4 Posterior recombination probability for full genome HIV-1 dataset after bias correction

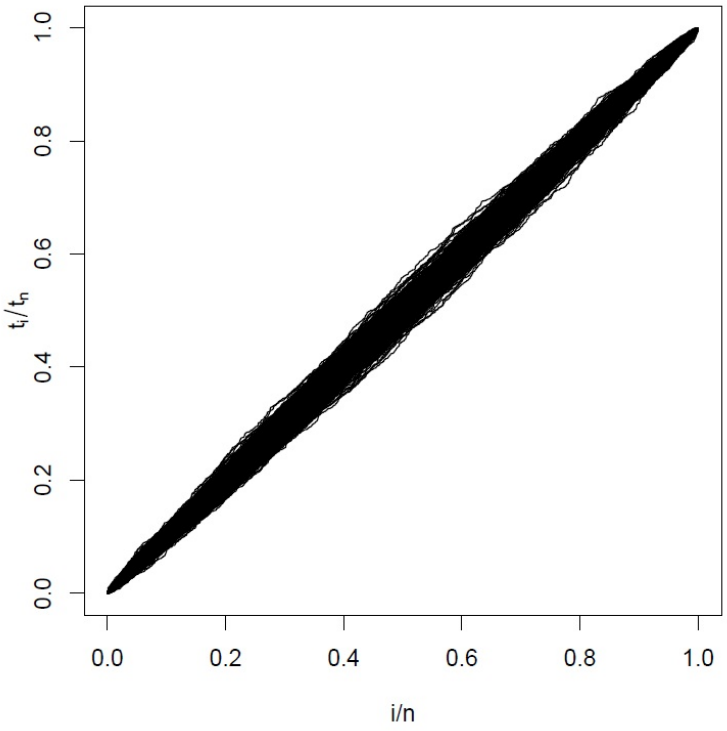



Figure 6.5 MCMC convergence diagnostics: SRQ plots for individual DMCP samplers

| Covariate | Mean coefficient | 95% BCI |
|---|---|---|
| *in vitro* hotspots | 0.1261 | $(-0.0131, 0.3587)$ |
| GC content (20) | 0.2760 | $(0.1817, 0.3805)$ |
| GC content (50) | 0.3185 | $(0.1932, 0.4404)$ |
| GC content (100) | 0.3489 | $(0.2120, 0.4595)$ |
| Entropy (sitewise) | $-2.5561$ | $(-2.9151, -1.9041)$ |
| Entropy (10) | 1.8577 | $(1.4276, 2.3050)$ |
| Entropy (20) | 1.9109 | $(1.505, 2.301)$ |
| Entropy (50) | 2.4460 | $(1.9691, 2.9172)$ |
| Deletions | 0.0081 | $(-0.0281, 0.0162)$ |
| Thermal Stability | $-0.2014$ | $(-0.3729, -0.0590)$ |
| G-rich stretches | 0.7361 | $(0.6269, 0.8428)$ |
| SHAPE reactivity | $-2.3934$ | $(-2.6328, -2.1549)$ |
| Pairing probability | 2.3896 | $(2.0642, 2.6857)$ |

Table 6.1 $\boldsymbol{\beta}$ coefficients inferred from full length HIV-1 genome analysis

# CHAPTER 7. ALTERNATE HIERARCHICAL GMRF PRIOR SPECIFICATION IN BINOMIAL RESPONSE MODELS

## 7.1 Introduction

In this chapter we motivate the use of the alternatives to the model described in earlier chapters. Briefly, in the hierarchical setup described hitherto, the upper level placed a GMRF prior on the population-level recombination probabilities $\boldsymbol{p}$. The lower level of the hierarchy inferred recombination breakpoints in individual sequences and these were communicated to the upper level through count data. This setup resulted in a binomial likelihood,

$$\pi(\boldsymbol{R} \mid \boldsymbol{p}) \propto \prod_{s=1}^{S} p_s^{R_s}(1-p_s)^{C_s-R_s} \tag{7.1}$$

where $R_s$ is the number of time site $s$ was a inferred to be a breakpoint at the lower level, $\boldsymbol{C}$ tracks the total number of times site $s$ was represented in the dataset and $S$ is the total length of the genome covered by the dataset. Placing a GMRF prior on the logit transformed recombination probability vector $\boldsymbol{\nu}$, Bayesian inference requires sampling from the density

$$\pi(\boldsymbol{\nu}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\nu}^T(\boldsymbol{Q}+\mathrm{diag}(\boldsymbol{c}))\boldsymbol{\nu} + \boldsymbol{b}^T\boldsymbol{\nu} + \sum f_s(\nu_s)\right) \tag{7.2}$$

where $f_s$ refers to the likelihood function at site $s$. With a binomial likelihood, Rue and Held (2005) suggest approximating the likelihood using a second-order Taylor se-

ries expansion and using the resultant GMRF approximation density as the proposal distribution in the Metropolis algorithm.

In the model discussed in this section, we apply an arcsin transformation on the count data such that the site-wise likelihood is asymptotically normal instead of binomial. This in turn results in the density (7.2) remaining a GMRF enabling direct sampling in a single Gibbs step as described in Chapter 5. This setup increases the speed of inference twofold. However, the proposed transformation method results in range restrictions on the inference that may limit its use to large datasets. We discuss the model setup , its merits and demerits in detail in the following sections.

## 7.2 Methods

We propose a toy example to illustrate the current model. At each site of the "genome" covered by the dataset the input data is in the form of normal random variates instead of sequence data. Each input vector was generated from a series of normal distributions thus creating changepoints in it. These changepoints are detected by the lower level of the hierarchical model. The upper level models the probability of a site to be a changepoint.

Suppose that we start with a dataset comprised of $K$ individual data vectors of lengths $L_1, L_2 \dots L_K$ respectively. Each data vector, $\boldsymbol{y_k} = (y_{k,1}, y_{k,2} \dots y_{k,L_k})$ is a set of $L_k$ observations that are piecewise i.i.d.. Further, each vector has $J$ changepoints such that observations in partition $j \in (1, 2, \dots J+1)$ are random variables from a $\mathcal{N}(\mu_j, \sigma_j^2)$ distribution.

### 7.2.1 Multiple Changepoint Model

At the lower level of the hierarchy our goal is to infer the number and location of the changepoints in individual data vectors. Denoting the model under $J$ changepoints

as $M_J$, the corresponding parameters are $\boldsymbol{\theta}_J = (\boldsymbol{\mu_J}, \boldsymbol{\sigma_J^2})$ where $\boldsymbol{\mu_J} = \{\mu_1, \mu_2 \dots \mu_{J+1}\}$ denotes the means of normal variates in the $J+1$ partitions and $\boldsymbol{\sigma_J^2}$ denotes the respective variances. The goal is to sample from the joint posterior

$$\pi(J, \boldsymbol{\theta}_J \mid \boldsymbol{y_k}) \propto \mathcal{L}(\boldsymbol{y_k} \mid J, \boldsymbol{\theta}_J)\pi(\boldsymbol{\theta}_J \mid J)P(J). \tag{7.3}$$

Conditional on the change points and means, the data are independent draws from normal distributions, so the likelihood is

$$\mathcal{L}(\boldsymbol{y_k} \mid J, \boldsymbol{\theta}_J) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma}\left(y_{k,s} - \mu(s)\right)^2\right] \tag{7.4}$$

where $\mu(s) = \mu_j$ for all $s$ belonging to partition $j$.

#### 7.2.1.1 Prior Specification

We assume the number of change points has a truncated Poisson prior

$$P(J) \propto \frac{e^{-\lambda}\lambda^J}{J!}, \qquad 0 \leq J \leq L_k,$$

where the constant of proportionality is the Poisson probability $P(k < n)$. Conditional on $J$, the locations of change points and means for the partitions created thereof, are independent. The locations are order statistics of discrete draws without replacement on $\{2, \dots, n\}$,

$$\pi(\boldsymbol{s} \mid J) = \frac{J!}{(n-1)\cdots(L_k - J)}.$$

We define normal priors on the means $\mu_j \in \boldsymbol{\mu_J} \sim \mathcal{N}(\mu_\mu, \sigma_\mu^2)$, specifically

$$\pi(\mu_j) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{1}{2\sigma_\mu^2}\left(\mu_j - \mu_\mu\right)^2\right].$$

where $\mu_\mu$ is the mean and $\sigma_\mu^2$ is the variance of the hyper-distribution. These are fixed at arbitrary values. The variance $\sigma_j^2$ for data in each partition is also fixed.

### 7.2.2 GMRF Prior on Changepoint Probabilities

At the upper level of the hierarchy, we analyze the entire dataset of $K$ data vectors. We summarize the information from the lower level inference in the form of two vectors $\boldsymbol{R}$ and $\boldsymbol{C}$. $R_s$ is the count of the number of times site $s$ was inferred to be a changepoint by the lower level analysis of the $K$ data vectors. $C_s$ is the number of trials/opportunities site $s$ had to be a changepoint. In other words, it is the number of times site $s$ is represented in the dataset.

#### 7.2.2.1 Arcsin data transformation

So far, the current model is identical in setup to the model described in previous chapters. However, instead of using the data directly in the form of counts $\boldsymbol{R}$ and $\boldsymbol{C}$,we now turn to the arcsin transformation of count data in order to use a normal likelihood model.

$$Z_s = \arcsin \sqrt{\frac{R_s}{C_s}}$$

When transformed thus, the distribution of $\boldsymbol{Z}$ is $\mathcal{N}(\arcsin(\sqrt{p_s}), 1/4C_s)$ asymptotically. Applying this transformation, the likelihood may now be expressed as,

$$\begin{aligned} \pi(Z_s \mid p_s) &\propto \mathcal{N}(\nu_s, 1/4C_s) \qquad (7.5) \\ \text{Implying ,} & \\ \pi(\boldsymbol{Z} \mid \boldsymbol{p}) &\propto \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{I/4C}) \end{aligned}$$

where $\nu_s = \arcsin(\sqrt{p_s})$. A simple GMRF with precision matrix as described in previous chapters is placed on the vector $\boldsymbol{\nu}$ with precision $\omega$.

### 7.2.3 Inference via MCMC Simulation

We sample from the joint posterior of all model parameters, (7.6), using MCMC simulation.

$$Pr(\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \ldots \boldsymbol{\theta_K}, \boldsymbol{\nu}, \omega \mid \boldsymbol{y_1}, \boldsymbol{y_2} \ldots \boldsymbol{y_K}) \propto \prod_{k=1}^{K} Pr(\boldsymbol{y_k} \mid \boldsymbol{\theta_k}) Pr(\boldsymbol{\theta_k}) \times Pr(\boldsymbol{Z} \mid \boldsymbol{\nu}) Pr(\boldsymbol{\nu} \mid \omega) Pr(\omega). \tag{7.6}$$

As before, we use a Metropolis-within-Gibbs scheme to update the model parameters in two major blocks. In the first block we simulate from the full conditional distribution of the lower level parameters

$$Pr(\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \ldots \boldsymbol{\theta_K} \mid \boldsymbol{\nu}, \omega, \boldsymbol{y_1}, \boldsymbol{y_2} \ldots \boldsymbol{y_K}) \propto \prod_{k=1}^{K} Pr(\theta_k | \boldsymbol{\nu}, y_k) = Pr(\boldsymbol{y_k} \mid \boldsymbol{\theta_k}) Pr(\boldsymbol{R_k} | \boldsymbol{\nu}) Pr(\boldsymbol{\psi_k}) \tag{7.7}$$

The second block of parameters consists of the $\boldsymbol{\nu}$ vector and the precision $\omega$.

$$Pr(\boldsymbol{\nu}, \omega) \propto Pr(\boldsymbol{Z} | \boldsymbol{\nu}) Pr(\boldsymbol{\nu} \mid \omega) Pr(\omega) \tag{7.8}$$

Unlike the scheme described in previous chapters, sampling from (7.8) may be achieved in a single Gibbs step when $\omega$ is fixed. The distribution we wish to sample from is of the general form as given in (7.2). Since the likelihood is normal, this remains a GMRF that we may directly sample from with methods described by Rue and Held (2005). When $\omega$ is not fixed it may be updated as described in Minin et al. (2007). However, the acceptance ratio should pertain only to $\omega$ in this case.

## 7.3 Results

### 7.3.1 Simulated Dataset

To test the working of current model we generated a simulated dataset of $K = 100$ data vectors each of length $L_1 = L_2 = \ldots = L_k = 100$ sites. For each dataset, the number of changepoints $J$ was fixed at four with locations fixed at positions $21, 41, 61$

and 81. Normal random variates within each of the five partitions created thereof, were generated from $\mathcal{N}(\mu_j, 1)$ distributions. Each $\mu_j$, in turn, was generated from a $\mathcal{N}(5, 1)$ distribution. Fig. 7.1 provides a schematic view of the simulation of a data vector.

We generated 11000 samples from the joint posterior, discarding the first 1000 as burn-in and sub-sampling the rest at every 10 samples to generate 1000 posterior samples. At the lower level, the sampler identified the 4 changepoints in all 100 data vectors accurately. Fig. 7.2 shows the posterior mean of the GMRF $\boldsymbol{\mu}$ with $\omega$ fixed at a value of 10.0.

MCMC convergence diagnostics were performed at the lower level, at the number of changepoints and at the upper level on the inferred vector $\boldsymbol{\nu}$ as described in the preceding chapter and the sampler was found to have converged at both levels (result not shown).

## 7.4 Discussion

We presented an alternative to the binomial likelihood representation of count data in hierarchical GMRF models. The arcsin transformation is advantageous in the case where a GMRF prior is applied to the upper level parameters since it specifies a normal likelihood. For normal response models, a sample may be drawn directly from the full conditional as described in (7.2) that remains a GMRF. When $\omega$ is fixed, the Gibbs algorithm may be employed to generate new MCMC samples. When $\omega$ is to be estimated, we use a Metropolis-within-Gibbs approach, updating $\boldsymbol{\nu}$ in a Gibbs step as described above followed by a Metropolis step for $\omega$.

Finding the GMRF approximation of a non-normal likelihood is computationally intensive and relates directly to the length $S$ of the region being analyzed. Doing away with this step increases the speed of inference twofold. The previous chapter discusses the advantages to adding a large number of sequences to the input dataset. Indeed

the model described in the previous chapter is also able to handle a large number of sequences. However, with the increase in speed afforded by the current model, even larger datasets may be analyzed in a reasonable length of time.

The arcsin transformation stabilizes the variance of the data. However, it restricts the range of the parameter to $(-\pi/2, \pi/2)$ which is further restricted to $(0, \pi/2)$ when dealing with probabilities. In a simple input data case such as the one presented in this example, this is not a problem. However, as the real probability goes closer to zero, as may be the case in sequence data, especially in regions that are conserved "cold spots" for recombination, the inference produces negative values that are outside the allowable range (data not shown).

Several solutions are currently being explored for this problem, including fixing a mean level and increasing the number of sequences. Once this problem has been addressed, the current model will be powerful in that it combines all the advantages of previous models with faster inference.

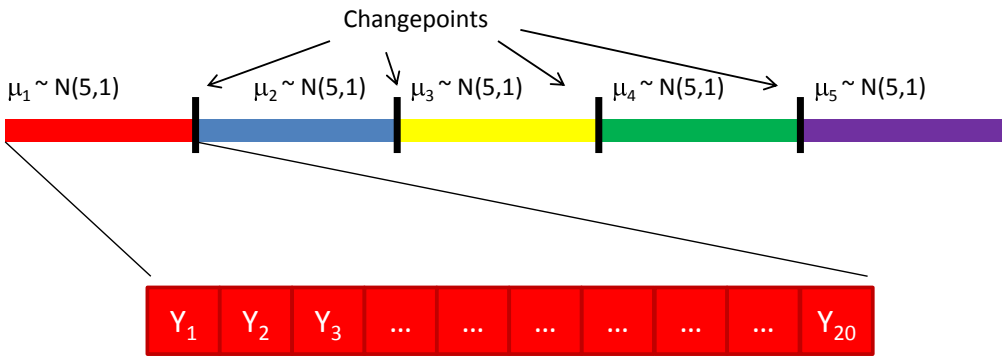



Figure 7.1 Schematic representation of data generation

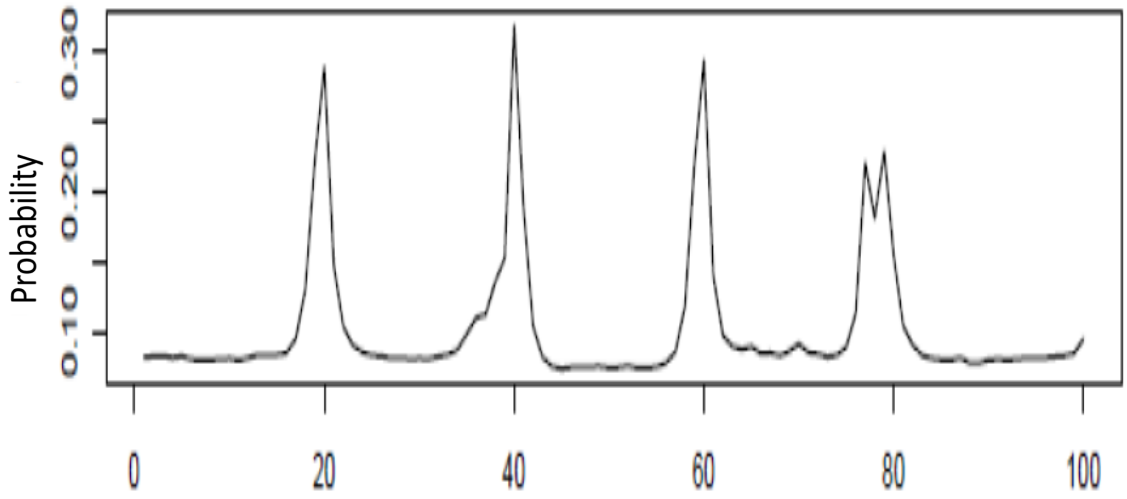



Figure 7.2 Posterior mean of $\mu$ in the hierarchical changepoint model

# CHAPTER 8. GENERAL CONCLUSIONS AND FUTURE WORK

The genetic diversity of HIV has long been a challenge in efforts of developing preventive therapies against the pathogen. Adding to the diversity are chimeric molecules produced by recombination. The genetic variants thus formed, along with the existing non-recombinant forms results in a wide range of host-pathogen interactions. Efficient clinical management of the disease necessitates, as a first step, the ability to identify the infecting genotype effectively. Further, a stepping stone towards effective therapeutic intervention of infection and subsequent progression of the disease is, veritably, better understanding of the mechanism of recombination. In this thesis, I described a rapid HIV genotyper based on supervised learning algorithms, a hierarchical model for simultaneous inference of spatial variation of recombination probabilities and covariates of interest and the application of this model to curated HIV datasets to gain insights into the mechanism of recombination.

## 8.1 HIV Genotyping

Application of machine learning algorithms to the problem of genotyping HIV sequences proved to be very successful. Compared to current genotyping tools, it is able to better classify complex recombinants. The success of the HIV genotyper can be augmented by providing many pre-made classifiers. Towards this, the short-term goal is to make the web tool a community-based effort, encouraging users to train classifiers

specific to sequences of their interest and share these with the community. Further, this genotyping method could find valuable application in the classification of bacterial sequences as well as classification of other retroviral sequences. In fact, when provided with a relevant training set, the genotyper may be able to classify retroviral sequences efficiently with the current feature sets. Bacterial sequences, on the other hand, may require more calibration in terms of optimized feature sets as well as the most suited supervised learning algorithm.

## 8.2 Hierarchical GMRF Model

We found, through analyses presented in this thesis, evidence supporting the hypothesis that propensity to secondary structures is directly correlated with high recombination rates. The framework provided by this model can be exploited to further dissect the molecular mechanism of recombination. In the future, I plan to extend the current model to include covariates at the level of the individual sequence. Such analyses can provide valuable insights into evolution of the virus. Covariates associated with recombination probabilities in such a set up will provide evidence for selection pressures acting on various regions of the virus. Further, epidemiological evidence can also be gathered through association of recombination rates with geographic location, time of infection and the interaction of these covariates with genotypes.

# BIBLIOGRAPHY

Andersen, E. S., Heeninga, R. E., Damgaard, C. K., Berkhout, B., and Kjems, J. (2003). Dimerization and template switching in the 5' untranslated region between various subtypes of Human Immunodeficiency Virus type 1. *J. Virol.*, 77:3020–3030.

Anderson, J. A., Teufel II, R. J., Yin, P. D., and Hu, W. S. (1998). Correlated template-switching events during minus-strand DNA synthesis: a mechanism for high negative interference during retroviral recombination. *J. Virol.*, 72:1186–1194.

Anderson, J. P., Rodrigo, A. G., Learn, G. H., Madan, A., Delahunty, C., Coon, M., Girard, M., Osmanov, S., Hood, L., and Mullins, J. (2000). Testing the hypothesis of a recombinant origin of Human Immunodeficiency Virus type 1 subtype E. *J. Virol.*, 74:10752–10765.

Arens, M. (1999). Methods for subtyping and molecular comparison of human viral genomes. *Clin. Microbiol. Rev.*, 12:612–626.

Baeten, J. M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R. S., Certain, L., Mandaliya, K., Jaoko, W., and Overbaugh, J. (2006). HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *J. Infect. Dis.*, 195(7):1177–1180.

Bagnarelli, P., Vecchi, M., Burighel, N., Bellanova, D., Menzo, S., Clementi, M., and Rossi, A. D. (2004). Genotypic and phenotypic correlates of the hiv type 1 env gene

evolution in infected children with discordant response to antiretroviral therapy. *AIDS Research and Human Retroviruses*, 20(12):1306–1313.

Baird, H. A., Galetto, R., Gao, Y., Simon-Loriere, E., Abreha, M., Archer, J., Fan, J., Robertson, D. L., Arts, E. J., and Negroni, M. (2006a). Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. *Nucleic Acids Res.*, 34:5203–5216.

Baird, H. A., Gao, Y., Galetto, R., Lalonde, M., Anthony, R. M., Giacomoni, V., Abreha, M., Destefano, J. J., Negroni, M., and Arts, E. J. (2006b). Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination. *Retrovirology*, 3:91.

Balakrishnan, M., Fay, P. J., and Bambara, R. A. (2001). The kissing hairpin sequence promotes recombination within the HIV-I 5' leader region. *J. Biol. Chem.*, 276:36482–35492.

Balakrishnan, M., Roques, B. P., Fay, P. J., and Bambara, R. A. (2003). Template dimerization promotes an acceptor invasion-induced transfer mechanism during human immunodeficiency virus type 1 minus-strand synthesis. *J. Virol.*, 77(8):4710–4721.

Bano, A. S., Sood, V., Neogi, U., Goel, N., Kuttiat, V. S., Wanchu, A., and Banerjea, A. C. (2009). Genetic and functional characterization of human immunodeficiency virus type 1 VprC variants from north india: presence of unique recombinants with mosaic genomes from B, C and D subtypes within the open reading frame of Vpr. *J Gen Virol.*, 90(11):2768–2776.

Bello, G., Eyer-Silva, W. A., Couto-Fernandez, J. C., Guimarales, M. L., Chequer-Fernandez, S. L., Teixeira, S. L. M., and Morgado, M. G. (2007). Demographic history of HIV-1 subtypes B and F in Brazil. *Infect. Genet. Evol.*, 7:263–270.

Berkhout, B., Vastenhouw, N. L., Klasens, B. I., and Huthoff, H. (2001). Structural features in the HIV-1 repeat region facilitate strand transfer during reverse transcription. *RNA*, 7:1097–1114.

Bernardinelli, L., C, P., Best, N. G., and Gilks, W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, 16:741–752.

Besag, J. E., York, J., and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43:1–59.

Best, N., Cowles, M., and Vines, K. (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.30. Technical report, MRC Biostatistics Unit, University of Cambridge.

Bhanja, P., Sengupta, S., Banerjee, D., Sarkar, K., Jana, S., and Chakrabarti, S. (2007). Detection of intersubtype recombinants with respect to env and nef genes of HIV-1 among female sex workers in Calcutta, India. *Virus Res*, 130(1):31–314.

Blackard, J. T., Renjifo, B., and Fawzi, W. e. a. (2001). HIV-1 LTR subtype and perinatal transmission. *Virology*, 287:261–265.

Bratt, G., Leandersson, A. C., Albert, J., Sandstrom, E., and Wahren, B. (1998). MT-2 tropism and CCR-5 genotype strongly influence disease progression in HIV-1-infected individuals. *AIDS*, 12:729–736.

Breiman, L. (2001). Random forests. Technical report, Department of Statistics, University of California, Berkeley.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Inc.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007). Human Immunodeficiency Virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *J. Virol.*, 81(19):10209–10219.

Burke, D. S. (1997). Recombination in HIV: An important viral evolutionary strategy. *Emerg. Infect. Dis.*, 3:253–259.

Butler, I. F., Pandrea, I., Marx, P. A., and Apetrie, C. (2007). Hiv genetic diversity: Biological and public health consequences. *Current HIV Research*, 5:23–45.

Chan, D. and Kim, P. (1998). HIV entry and its inhibition. *Cell*, 93(5):681–684.

Charpentier, C., Nora, T., Tenaillon, O., Clavel, F., and Hance, A. J. (2006). Extensive recombination among Human Immunodeficiency Virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol.*, 80:2472–2482.

Chawla, N. V. (2006). Data mining for imbalanced datasets: An overview. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer US.

Chawla, N. V., Bowyer, K. W., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling TEchnique. *J. Artif. Intell. Res.*, 16:321–357.

Chen, Y., Balakrishnan, M., Roques, B. P., Fay, P. J., and Bambara, R. A. (2003). Mechanism of minus strand strong stop transfer in hiv-1 reverse transcription. *J. Biol. Chem.*, 278:8006–8017.

Chib, S. and Greenberg, E. (1995). Understanding the metropolishastings algorithm. *American Statistician*, 49(4):327–335.

Chin, M. P. S., Chen, J., Nikolaitchik, O. A., and Hu, W.-S. (2007). Molecular determinants of HIV-1 intersubtype recombination potential. *Virology*, 363:437–446.

Chipman, H., George, E. I., and McCulloch, R. E. (2008). BART: Bayesian additive regression trees. Technical report, Department of Mathematics and Statistics, Acadia University, Canada.

Chohan, B., Lavreys, L., Rainwater, S. M., and Overbaugh, J. (2005). Evidence for frequent reinfection with human immunodeficiency virus type 1 of a different subtype. *J. Virol.*, 79(16):10701–10708.

Clapham, P. R. and McKnight, A. (2001). HIV-1 receptors and cell tropism. *Br Med Bull*, 58(4):43–59.

Clark, S. A., Calef, C., and Mellors, J. W. (2005). Mutations in retroviral genes associated with drug resistance. In Leitner, T., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J. W., Wolinsky, S., and Korber, B., editors, *HIV Sequence Compendium*, pages 80–174. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

Coffin, J. M. (1979). Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J. gen. Virol.*, 42:1–26.

Coffin, J. M., Hughes, S. H., and Varmus, H. E., editors (1997). *Retroviruses*. Cold Spring Harbor Laboratory Press.

Collman, R., Hassan, N. F., Walker, R., Godfrey, B., Cutilli, J., Hastings, J. C., Friedman, H., Douglas, S. D., and Nathanson, N. (1989). Infection of monocyte-derived macrophages with human immunodeficiency virus type 1 (HIV-1). Monocyte-tropic and lymphocyte-tropic strains of HIV-1 show distinctive patterns of replication in a panel of cell types. *J. Exp. Med.*, 170:1149–1163.

Colson, P., Solas, C., Moreau, J., Motte, A., Henry, M., and Tamalet, C. (2007). Impaired quantification of plasma HIV-1 RNA with a commercialized real-time PCR assay in a couple of HIV-1-infected individuals. *J. Clin. Virol.*, 39(3):226–229.

Cornelissen, M., Mulder-Kampinga, G., Veenstra, J., Zorgdrager, F., Kuiken, C., Hartman, S., and et al. (1995). Syncytium-inducing (SI) phenotype suppression at seroconversion after intramuscular inoculation of a non-syncytium-inducing/si phenotypically mixed human immunodeficiency virus population. *J. Virol.*, 69:810–818.

Cornelissen, M., van den Burg, R., Zorgdrager, F., Lukashov, V., and Goudsmit, J. (1997). Pol gene diversity of five Human Immunodeficiency Virus type 1 subtypes: Evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns, and common ancestry for subtypes B and D. *J. Virol.*, 71:6348–6358.

Cuevas, M. T., Ruibal, I., Villahermosa, M. L., Diaz, H., Delgado, E., Parga, H. V., Perez-Alvarez, L., de Armas, M. B., Cuevas, L., Medrano, L., Noa, E., Osmanov, S., Najera, R., and Thomson, M. M. (2002). High HIV-1 genetic diversity in Cuba. *AIDS*, 16:1643–1653.

de Oliveira, T., Deforche, K., Cassol, S., Salminem, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M. J., van de Vijver, D. A., Boucher, C. A., Camacho, R., and Vandamme, A. M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21(19):3797–3800.

Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA*, 106(1):97–102.

Derebail, S. S., Heath, M. J., and DeStefano, J. J. (2003). Evidence for the differential effects of nucleocapsid protein on strand transfer in various regions of the HIV genome. *J. Biol. Chem.*, 278:15702–15712.

Desport, M., editor (2010). *Lentiviruses and Macrophages: Molecular and Cellular Interactions*. Caister Academic Press.

DeStefano, J. J., Wu, W., Seehra, J., McCoy, J., Laston, D., Albone, E., Fay, P. J., and Bambara, R. A. (1996). Characterization of an rnase h deficient mutant of human immunodeficiency virus-1 reverse transcriptase having an aspartate to asparagine change at position 498. *Biochim. Biophys. Acta*, 1219:380–388.

Diaz, L. and DeStefano, J. J. (1996). Strand transfer is enhanced by mismatched nucleotides at the 3' primer terminus: a possible link between hiv reverse transcriptase fidelity and recombination. *Nucleic Acids Res*, 24:3086–3092.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 39:103–137.

Dorman, K. S., Kaplan, A. H., and Sinsheimer, J. S. (2002). Bootstrap confidence levels for HIV-1 recombination. *J Mol Evol*, 54(2):200–209.

Douek, D. C., Roederer, M., and Koup, R. A. (2009). Emerging concepts in the immunopathogenesis of AIDS. *Annu. Rev. Med.*, 60:471–84.

Douglas, N., Knight, A. I., Hayhurst, A., Barrett, W. Y., Kevany, M. J., and Daniels, R. S. (1996). An efficient method for the rescue and analysis of functional HIV-1 *env* genes: Evidence for recombination in the vicinity of the *tat*/*rev* splice site. *AIDS*, 10:39–46.

Dykes, C., Balakrishnan, M., Planelles, V., Zhu, Y., Bambara, R. A., and Demeter, L. M. (2004). Identification of a preferred region for recombination and mutation in HIV-1 gag. *Virology*, 326(2):262–279.

Esbjornsson, J., Mansson, F., Martinez-Arias, W., Vincic, E., Biague, A. J., da Silva, Z. J., Fenyo, E. M., Norrgren, H., and Medstrand, P. (2010). Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease–indication of an evolving epidemic in West Africa. *Retrovirology*, 7:23.

Fan, J., Negroni, M., and Robertson, D. L. (2007). The distribution of HIV-1 recombination breakpoints. *Infect Genet Evol*, 7(6):717–723.

Fang, F., Ding, J., N., M. V., Suchard, M. A., and Dorman, K. S. (2007). cBrother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics*, 23(4):507–508.

Fang, G., Weiser, B., Kuiken, C., Philpott, S. M., Rowland-Jones, S., Plummer, F., Kimani, J., Shi, B., Kaul, R., Bwayo, J., Anzala, O., and Burger, H. (2004). Recombination following superinfection by HIV-1. *AIDS*, 18:153–159.

Frange, P., Galimand, J., Vidal, N., Goujard, C., Deveau, C., Souala, F., Peeters, M., Meyer, L., Rouzioux, C., and Chaix, M. L. (2008). New and old complex recombinant HIV-1 strains among patients with primary infection in 1996-2006 in France: the French ANRS CO06 primo cohort study. *Retrovirology*, 5:69.

Gale, C. V., Myers, R., Tedder, R. S., Williams, I. G., and Kellam, P. (2004). Development of a novel Human Immunodeficiency Virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London. *AIDS Res. Hum. Retrov.*, 20:457–464.

Galetto, R., Giacomoni, V., Veron, M., and Negroni, M. (2006). Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *J. Biol. Chem.*, 281(8):2711–2720.

Galetto, R., Moumen, A., Giacomoni, V., Veron, M., Charneau, P., and Negroni, M. (2004). The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J. Biol. Chem.*, 279:36625–36632.

Galetto, R. and Negroni, M. (2005). Mechanistic features of recombination in HIV. *AIDS Rev.*, 7:92–102.

Galli., A., Lai, A., Corvasce, S., Saladini, F., Riva, C., Deho, L., Caramma, I., Franzetti, M., Romano, L., Galli, M., Zazzi, M., and C., B. (2008). Recombination analysis and structure prediction show correlation between breakpoint clusters and rna hairpins in the pol gene of human immunodeficiency virus type 1 unique recombinant forms. *J Gen Virol*, 89(12):3119–3125.

Gao, L., Balakrishnan, M., Roques, B. P., and Bambara, R. A. (2007). Insights into the multiple roles of pausing in HIV-1 reverse transcriptase-promoted strand transfers. *J. Biol. Chem.*, 282:6222–6231.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., and Korber, B. (2002). Diversity Considerations in HIV-1 Vaccine Selection. *Science*, 296(5577):2354–2360.

Gelman, A., Carlin, J. B., and Stern, H. S. (2004). *Bayesian Data analysis*. Chapman and Hall/CRC Texts in Statistical Science.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7:457–472.

Gendelman, H. E., Orenstein, J. M., Martin, M. A., Ferrua, C., Mitra, R., Phipps, T., Wahl, L. A., Lane, H. C., Fauci, A. S., and Burke, D. S. (1988). Efficient isolation and propagation of human immunodeficiency virus on recombinant colony-stimulating factor 1-treated monocytes. *J. Exp. Med.*, 169:1428–1441.

Geretti, A. (2006). HIV-1 subtypes: Epidemiology and significance for HIV management. *Curr. Opin. Infect. Dis.*, 19(1):1–7.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics*, pages 169–193. Oxford University Press.

Gifford, R., de Oliveira, T., Rambaut, A., Myers, R. E., Gale, C. V., Dunn, D., Shafer, R., Vandamme, A.-M. nd Kellam, P., Pillay, D., and on HIV Drug Resistance, U. K. C. G. (2006). Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. *AIDS*, 20:1521–1529.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, 44(4):455–472.

Goodenow, M. M. and Collman, R. G. (2006). HIV-1 coreceptor preference is distinct from target cell tropism: a dual-parameter nomenclature to define viral phenotypes. *J Leukoc Biol*, 80(5):965–972.

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

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732.

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):160–174.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosciences*, 98(2):185 – 200.

Hill, M., Tachedjian, G., and Mak, J. (2005). The packaging and maturation of the HIV-1 pol proteins. *Curr HIV Res*, 3(1):73–85.

HIV-Database (2010). Los alamos HIV sequence databases. http://www.hiv.lanl.gov/.

Holguin, A., Lopez, M., and Soriano, V. (2008a). Reliability of rapid subtyping tools compared to phylogenetic analysis for characterization of HIV-1 non-B subtypes and recombinant forms. *J. Clin. Microbiol.*, 46(12):3896–3899.

Holguin, A., Lospitao, E., Lopez, M., Ramirez de Arellano, E., Pena, M. J., del Romero, J., Martin, C., and Soriano, V. (2008b). Genetic characterization of complex inter-recombinant HIV-1 strains circulating in Spain and reliability of distinct rapid subtyping tools. *J. Med. Virol.*, 80:383–391.

Hu, D. J., Subbarao, S., Vanichseni, S., Mock, P. A., Ramos, A., Nguyen, L., Chaowanachan, T., van Griensven, F., Choopanya, K., Mastro, T. D., and Tappero, J. W. (2005). Frequency of HIV-1 dual subtype infections, including intersubtype superinfections, among injection drug users in Bangkok, Thailand. *AIDS*, 19:303–308.

Hu, W. S. and Temin, H. M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *P. Natl. Acad. Sci. USA*, 87:1556–1560.

Husmeier, D. and McGuire, G. (2002). Detecting recombination with MCMC. *Bioinformatics*, 18(suppl 1):S345–353.

Husmeier, D. and McGuire, G. (2003). Detecting recombination in 4-taxa dna sequence alignments with bayesian hidden markov models and markov chain monte carlo. *Mol Biol Evol*, 20(3):315–337.

Husmeier, D. and Wright, F. (2001). Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics*, 17(suppl 1):S123–131.

Iweala, O. I. (2004). HIV diagnostic tests: An overview. *Contraception*, 70(2):141–147.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D., and Dougherty, J. P. (2000). High rate of recombination throughout the Human Immunodeficiency Virus type 1 genome. *J. Virol.*, 74:1234–1240.

Kahn, J. O. and Walker, B. D. (1998). Acute human immunodeficiency virus type 1 infection. *N Engl J Med*, 339:33–39.

Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., and Lyagoba, F. (2002). Effect of Human Immunodeficiency Virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. *J. Infect. Dis.*, 185:1244–50.

Kilmarx, P. (2008). Acquired immunodeficiency syndrome. In Heyman, D. L., editor, *Control of communicable diseases manual*, pages 1–10. APHA Press, 19 edition.

Kim, J. K., Palaniappan, C., Wu, W., Fay, P. J., and Bambara, R. A. (1997). Evidence for a unique mechanism of strand transfer from the transactivation response region of HIV-1. *J. Biol. Chem.*, 272:16769–16777.

Kinomoto, M., Yokoyama, M., Sato, H., Asato, K., Kurata, T., Ikuta, K., Sata, T., and Tokunaga, K. (2005). Amino Acid 36 in the Human Immunodeficiency Virus Type 1 gp41 Ectodomain Controls Fusogenic Activity: Implications for the Molecular Mechanism of Viral Escape from a Fusion Inhibitor. *J. Virol.*, 79(10):5996–6004.

Klarmann, G. J., Schauber, C. A., and Preston, B. D. (1993). Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 sequences in vitro. *J. Biol. Chem.*, 268:9793–9802.

Korber, B., Foley, B. T., Kuiken, C., Pillai, S. K., and Sodroski, J. G. (1998). Numbering positions in hiv relative to hxb2cg. In Korber, B., Kuiken, C. L., Foley, B., Hahn, B., McCutchan, F., Mellors, J. W., and Sodroski, J., editors, *Human Retroviruses and AIDS Compendium*, pages III–102–III–111. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

Korber, B. and Gnanakaran, S. (2009). The implications of patterns in HIV diversity for neutralizing antibody induction and susceptibility. *Curr Opin HIV AIDS*, 4(5):408–417.

Kuiken, C., Foley, B., Leitner, T., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., Wolinsky, A., and Korber, B., editors (2010). *HIV Sequence Compensium 2010*. Los Alamos National Laboratory.

Lanciault, C. and Champoux, J. J. (2006). Pausing during reverse transcription increases the rate of retroviral recombination. *J. Virol.*, 80:2483–2494.

Levy, D. N., Aldrovandi, G. M., Kutsch, O., and Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. *P. Natl. Acad. Sci. USA*, 101(12):4204–4209.

Liitsola, K., Tashkinova, I., Laukkanen, T., Korovina, G., Smolskaja, T., Momot, O., Mashkilleyson, N., Chaplinskas, S., Brummer-Korvenkontio, H., Vanhatalo, J.,

Leinikki, P., and Salminen, M. O. (1998). HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS*, 12:1907–19.

MacNeil, A., Sankale, J. L., Meloni, S. T., Sarr, A. D., Mboup, S., and Kanki, P. (2007). Long-term intrapatient viral evolution during HIV-2 infection. *J Infect Dis*, 195(5):726–733.

Madani, N., Perdigoto, A. L., Srinivasan, K., Cox, J. M., Chruma, J. J., LaLonde, J., Head, M., Smith, A. B. r., and Sodroski, J. G. (2004). Localized changes in the gp120 envelope glycoprotein confer resistance to Human Immunodeficiency Virus entry inhibitors BMS-806 and 155. *J. Virol.*, 78(7):3742–3752.

Magiorkinis, G., Paraskevis, D., Vandamme, A.-M., Magiorkinis, E., Vana Sypsa, V., and Hatzakis, A. (2003). In vivo characteristics of HIV-1 intersubtype recombination: Determination of hot spots and correlation with sequence similarity. *J. Gen. Virol.*, 84(10):2715–2722.

McCutchan, F. E. (2006). Global epidemiology of HIV. *J. Med. Virol.*, 78:S1–S7.

McGuire, G., Wright, F., and Prentice, M. J. (1997). A graphical method for detecting recombination in phylogenetic data sets. *Mol Biol Evol*, 14(11):1125–1131.

Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13):3034–3042.

Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. (2007). Phylogenetic mapping of recombination hot-spots in HIV via spatially smoothed change-point processes. *Genetics*, 175:1773–1785.

Moore, M. D., Fu, W., Nikolaitchik, O., Chen, J., Ptak, R. G., and Hu, W.-S. (2007). Dimer initiation signal of HIV-1: its role in partner selection during RNA copackaging and its effects on recombination. *J. Virol.*, 81:4002–4011.

Moumen, A., Polomack, L., Roques, B., Buc, H., and Negroni, M. (2001). The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res.*, 29:3814–3821.

Moumen, A., Polomack, L., Unge, T., Veron, M., Buc, H., and Negroni, M. (2003). Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA. *J. Biol. Chem.*, 278:15973–15982.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in markov chain samplers. *J. Am. Stat. Assoc*, 90:233–241.

Negroni, M. and Buc, H. (1999). Recombination during reverse transcription: an evaluation of the role of the nucleocapsid protein. *J. Mol. Biol.*, 9286:15–31.

Negroni, M. and Buc, H. (2000). Copy-choice recombination by reverse transcriptases: reshuffling of genetic markers mediated by RNA chaperones. *P. Natl. Acad. Sci. USA*, 97(12):6385–6390.

Negroni, M. and Buc, H. (2001). Mechanisms of retroviral recombinaiton. *Annu. Rev. Genet.*, 35:275–302.

Neilson, J. R., John, G. C., Carr, J. K., Lewis, P., Kreiss, J. K., and Jackson, S. (1999). Subtypes of Human Immunodeficiency Virus type 1 and disease stage among women in nairobi, kenya. *J. Virol.*, 73:4393–4403.

Njai, H. F., Gali, Y., Vanham, G., Clybergh, C., Jennes, W., Vidal, N., Butel, C., Mpoudi-Ngolle, E., Peeters, M., and Ariën, K. K. (2006). The predominance of HIV-1

circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. *Retrovirology*, 3:40.

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F., and Hance, A. J. (2007). Contribution of recombination to the evolution of HIV expressing resistance to antiretroviral treatment. *J. Virol.*, 81(14):7280–7288.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217.

Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P., and Hein, J. (2004). A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nuc Acid Res*, 32:4925–4936.

Peeters, M. (2000). Recombinant HIV sequences: Their role in the global epidemic. In Kuiken, C., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Mullins, J., Wolinsky, S., and Korber, B., editors, *HIV Sequence Compendium*, pages I39–I54. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

Peeters, M., Vincent, R., Perret, J. L., Lasky, M., Patrel, D., and Liegeois, F. (1999). Evidence for differences in MT2 cell tropism according to genetic subtypes of HIV-1: syncytium-inducing variants seem rare among subtype C HIV-1 viruses. *J. Acq. Imm. Def.*, 20:115–121.

Philpott, S. M. (2003). Hiv-1 coreceptor usage, transmission, and disease progression. *Curr HIV Res*, 1(2):217–227.

Piantadosi, A., Chohan, B., Chohan, V., McClelland, R. S., and Overbaugh, J. (2007). Chronic HIV-1 infection frequently fails to protect against superinfection. *PLOS Pathog*, 3(11):e177.

Ping, L. H., Nelson, J. A., Hoffman, I. F., Schock, J., Lamers, S. L., Goodman, M., Vernazza, P., Kazembe, P., Maida, M., Zimba, D., Goodenow, M. M., Eron, J. J. J., Fiscus, S. A., Cohen, M. S., and Swanstrom, R. (1999). Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. *J Virol*, 73:6271–6281.

Piyasirisilp, S., McCutchan, F. E., Carr, J. K., Sanders-Buell, E., Liu, W., Chen, J., Wagner, R., Wolf, H., Shao, Y., Lai, S., Beyrer, C., and Yu, X. F. (2000). A recent outbreak of Human Immunodeficiency Virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. *J. Virol.*, 74:11286–11295.

Plantier, J. C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemee, V., Damond, F., Robertson, D. L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nat Med*, 15(8):871–872.

Posada, D. and Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *P. Natl. Acad. Sci. USA*, 98:13757–13762.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *AAAI 2000 Workshop on Imbalanced Data Sets*.

Quinones-Mateu, M. E., Gao, Y., Ball, S. C., Marozsan, A. J., Abraha, A., and Arts, E. J. (2002). In vitro intersubtype recombinants of Human Immunodeficiency Virus type 1: comparison to recent and circulating in vivo recombinant forms. *J. Virol.*, 76:9600–9613.

Rajaram, M. L. and Dorman, K. S. (2009). Rapid genotyping of hiv-1 sequences using supervised learning algorithms. In *Proceedings of the International Conference on*

*Bioinformatics & Computational Biology, BIOCOMP 2009*, pages 334–339. CSREA Press.

Rajaram, M. L., Minin, V. N., Suchard, M. A., and Dorman, K. S. (2007). Hot and cold: Spatial fluctuation in hiv-1 recombination rates. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007*, pages 707–714. IEEE.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of hiv evolution. *Nat Rev Gen*, 5:52–61.

Ramirez, B. C., Simon-Loriere, E., Galetto, R., and Negroni, M. (2008). Implications of recombination for HIV diversity. *Virus Res*, 134(1-2):64–73.

Renjifo, B., Gilbert, P., Chaplin, B., Vannberg, F., Mwakagile, D., Msamanga, G., Hunter, D., Fawzi, W., and Essex, M. (1999). Emerging recombinant Human Immunodeficiency Viruses: uneven representation of the envelope V3 region. *AIDS*, 13:1613–1621.

Resnik, P. and Hardisty, E. (2009). Gibbs sampling for the uninitiated. Technical report, University of Maryland College Park, MD.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science*, 288(5463):55–56.

Robertson, D. L., Hahn, B. H., and Sharp, P. M. (1995). Recombination in aids viruses. *J Mol Evol*, 40(3):249–259.

Roda, R. H., Balakrishnan, M., Hanson, M. N., Wohrl, B. M., Le Grice, S. F., Roques, B. P., Gorelick, R. J., and Bambara, R. A. (2003). Role of the reverse transcriptase, nucleocapsid protein, and template structure in the two-step transfer mechanism in retroviral recombination. *J. Biol. Chem.*, 278:31536–31546.

Roda, R. H., Balakrishnan, M., Kim, J. K., Roques, B. P., Fay, P. J., and Bambara, R. A. (2002a). Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. *J. Biol. Chem*, 277:46900–46911.

Roda, R. H., Balakrishnan, M., Kim, J. K., Roques, B. P., Fay, P. J., and Bambara, R. A. (2002b). Strand Transfer Occurs in Retroviruses by a Pause-initiated Two-step Mechanism. *J Bio Chemistry*, 277(49):46900–46911.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., and Tatusova, T. (2004). A web-based genotyping resource for viral sequences. *Nucleic Acid Res.*, 32:W654–W659.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC.

Sabino, E. C., Shpaer, E. G., G., M. M., T., K. B., S., D. R., and et al, B. V. (1994). Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from brazil. *J. Virol*, 68:6340–6346.

Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of hiv type 1 by bootscanning. *AIDS Res Hum Retroviruses*, 11(11):1423–1425.

Salminen, M. O., Carr, J. K., Robertson, D. L., Hegerich, P., Gotte, D., Koch, C., Sanders-Buell, E., Gao, F., Sharp, P. M., Hahn, B. H., Burke, D. S., and McCutchan, F. E. (1997). Evolution and probable transmission of intersubtype recombinant Human Immunodeficiency Virus type 1 in a Zambian couple. *J. Virol.*, 71:2647–2655.

Sanders-Buell, E., Bose, M., Nasir, A., Todd, C. S., Stanekzai, M. R., Tovanabutra, S., Scott, P. T., Strathdee, S. A., Tjaden, J., and Michael N. L., McCutchan, F. E. (2010). Distinct circulating recombinant HIV-1 strains among injecting drug users and sex workers in afghanistan. *AIDS Res Hum Retroviruses*, 26(5):605–608.

Shen, W., Gao, L., Balakrishnan, M., and Bambara, R. A. (2009). A recombination hot spot in HIV-1 contains guanosine runs that can form a G-quartet structure and promote strand transfer in vitro. *J Biol Chem*, 284(49):33883–33893.

Smith, D., Richman, D., and Little, S. (2005). Hiv superinfection. *J Inf Dis*, 192(3):438–444.

Soto-Ramirez, L. E., Renjifo, B., McLane, M. F., Marlink, R., O.Hara, C., and Sutthent, R. (1996). HIV-1 langerhans. cell tropism associated with heterosexual transmission of HIV. *Science*, 271:1291–1293.

Spira, S., Wainberg, M. A., Loemba, H., Turner, D., and Brenner, B. G. (2003). Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. *J. Antimicrob. Chemoth.*, 51(2):229–240.

Suchard, M. A., Weiss, R. E., Dorman, K. S., and Sinsheimer, J. S. (2002). Oh brother, where art thou? a bayes factor test for recombination with uncertain heritage. *Syst. Biol.*, 51(5):715728.

Suchard, M. A., Weiss, R. E., Dorman, K. S., and Sinsheimer, J. S. (2003). Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *J. Am. Stat. Assoc.*, 98:427–437.

Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., and Sasaki, M. (1995). Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, 34(35):11211–11216.

Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika*, 86(2):341–350.

Suo, Z. and Johnson, K. A. (1997). Effect of rna secondary structure on the kinetics of dna synthesis catalyzed by hiv-1 reverse transcriptase. *Biochemistry*, 36:12459–12467.

Swanson, P., de Mendoza, C., Joshi, Y., Golden, A., Hodinka, R. L., Soriano, V., and et. al (2005). Impact of Human Immunodeficiency Virus type 1 (HIV-1) genetic diversity on performance of four commercial viral load assays: LCx HIV RNA Quantitative, AMPLICOR HIV-1 MONITOR v1.5, VERSANT HIV-1 RNA 3.0, and NucliSens HIV-1 QT. *J. Clin. Microbiol.*, 43(8):3860–3868.

Takebe, Y., Motomura, K., Tatsumi, M., Lwin, H. H., Zaw, M., and Kusagawa, S. (2003). High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: geographical hot spot of extensive recombination. *AIDS*, 17:2077–2087.

Tee, K. K., Saw, T. L., Pon, C. K., Kamarulzaman, A., and Ng, K. P. (2005). The evolving molecular epidemiology of HIV type 1 among injecting drug users (IDUs) in Malaysia. *AIDS Res. Hum. Retrov.*, 21:1046–1050.

Temin, H. M. (1991). Sex and recombination in retroviruses. *Trends Genet.*, 7:71–74.

Templeton, A. R., Kramer, M. G., Jarvis, J., Kowalski, J., Gange, S., Schneider, M. F., Shao, Q., Zhang, G. W., Yeh, M. F., Tsai, H. L., Zhang, H., and Markham, R. B. (2009). Multiple-infection and recombination in HIV-1 within a longitudinal cohort of women. *Retrovirology*, 6:54.

Thomson, M. M., Sierra, M., Tanuri, A., May, S., Casado, G., Manjan, N., and Najera, R. (2004). Analysis of near full-length genome sequences of HIV type 1 BF intersubtype recombinant viruses from Brazil reveals their independent origins and their lack of relationship to CRF12_BF. *AIDS Res. Hum. Retrov.*, 20:1126–1133.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann Stat*, 22:1701–1762.

Toni, T. Adja-Toura, C., Vidal, N., Minga, A., Huet, C., Borger, M.-Y., Recordon-Pinson, P., Masquelier, B., Nolan, M., Nkengasong, J., Fluery, H. J., Delaporte, E., and Peeters, M. (2005). Presence of CRF09_cpx and complex CRF02_AG/CRF09_cpx recombinant HIV type 1 strains in Côte d'Ivoire, West Africa. *AIDS Res. Hum. Retrov.*, 21:667–672.

Tovanabutra, S., Beyrer, C., Sakkhachornphop, S., Razak, M. H., Ramos, G. L., Vongchak, T., Rungruengthanakit, K., Saokhieo, P., Tejafong, K., Kim, B., De Souza, M., Robb, M. L., Birx, D. L., Jittiwutikarn, J., Suriyanon, V., Celentano, D. D., and McCutchan, F. E. (2004). The changing molecular epidemiology of HIV type 1 among northern thai drug users, 1999 to 2002. *AIDS Res. Hum. Retrov.*, 20(5):465–475.

Tscherning, C., Alaeus, A., Fredriksson, R., Bjorndal, A., Deng, H., Littman, D. R., Fenyo, E. M., and Albert, J. (1998). Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. *Virology*, 241:181–188.

UNAIDS (2006). Overview of the aids epidemic. Technical report, Joint United Nations Programmer of HIV/AIDS.

Vergne, L., Peeters, M., Mpoudi-Ngole, E., Bourgeois, A., Liegeois, F., Toure-Kane, C., Mboup, S., Mulanga-Kabeya, C., Saman, E., Jourdan, J., Reynes, J., and Delaporte, E. (2000). Genetic Diversity of Protease and Reverse Transcriptase Sequences in Non-Subtype-B Human Immunodeficiency Virus Type 1 Strains: Evidence of Many Minor Drug Resistance Mutations in Treatment-Naive Patients. *J. Clin. Microbiol.*, 38(11):3919–3925.

Vogt, P. K. (1997). Historical introduction to the general properties of retroviruses. In Coffin, J. M., Hughes, S. H., and Varmus, H. E., editors, *Retroviruses*, pages 1–25. Cold Spring Harbor Laboratory Press.

Wain-Hobson, S. (1993). The fastest genome evolution ever described: HIV variation in situ. *Curr. Opin. Genet. Dev.*, 3:878–883.

Wang, B., Lau, K. A., Ong, L. Y., Shah, M., Steain, M. C., Foley, B., Dwyer, D. E., Chew, C. B., Kamarulzaman, A., Ng, K. P., and Saksena, N. K. (2007). Complex patterns of the HIV-1 epidemic in kuala lumpur, malaysia: evidence for expansion of circulating recombinant form CRF33_01B and detection of multiple other recombinants. *Virology*, 367:288–297.

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W. J., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009). Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716.

Weiss, R. A. (1993). How does HIV cause AIDS? *Science*, 260(5112):1273–1279.

WHO (2007). Towards universal access: Scaling up priority HIV/AIDS interventions in the health sector : Progress report, April 2007. Technical report, World Health Organization, UNAIDS, UNICEF.

Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc*, 1(3):1610–1616.

Worobey, M. and Holmes, E. C. (1999). Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.*, 80:2535–2543.

Wu, W., Blumberg, B. M., Fay, P. J., and Bambara, R. A. (1995). Strand transfer mediated by Human Immunodeficiency Virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation. *J. Biol. Chem.*, 270:325–332.

Wu, X., Cai, Z., Wan, X.-F., Hoang, T., Goebel, R., and Lin, G. (2007). Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*, 23(14):1744–1752.

Wyatt, R. and Sodroski, J. (1998). The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, 280(5271):1884–1888.

Yu, X. F., Wang, Z., Beyrer, C., Celentano, D. D., Khamboonruang, C., Allen, E., and Nelson, K. (1995). Phenotypic and genotypic characteristics of human immunodeficiency virus type 1 from patients with AIDS in northern Thailand. *J Virol*, 69:4649–4655.

Zhang, C. Y., Wei, J. F., and H., H. S. (2005). The key role for local base order in the generation of multiple forms of China HIV-1 B'/C intersubtype recombinants. *BMC Evol. Biol.*, 5.

Zhang, J., Tang, T. A., Li, T., Ma, Y., and Sapp, C. M. (2000). Most retroviral recombinations occur during minus-strand dna synthesis. *J. Virol.*, 74:2313–2322.

Zhang, J. L. and Hardle, W. (2008). The Bayesian additive classification tree applied to credit risk modelling. Technical report, Humboldt University Berlin, Germany.

Zhu, T., Wang, N., Carr, A., Nam, D. S., Moor-Jankowski, R., Cooper, D. A., and Ho, D. D. (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):3098–3107.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. (2002). Human Immunodeficiency Virus type 1 recombination: rate, fidelity, and putative hot spots. *J. Virol.*, 76:11273–11282.