Hot and Cold: Spatial Fluctuation in HIV-1 Recombination Rates

M. L. Rajaram
Bioinformatics & Computational Biology Program
Iowa State University
Ames, IA 50011
mishar@iastate.edu

V. N. Minin
Department of Statistics
University of Washington
Seattle, WA 98195
vminin@stat.washington.edu

M. A. Suchard
Departments of Biomathematics, Human Genetics, and Biostatistics
University of California, Los Angeles
Los Angeles, CA 90095
msuchard@ucla.edu

K. S. Dorman
Departments of Statistics and Genetics, Development & Cell Biology
Iowa State University
Ames, IA 50011
kdorman@iastate.edu

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.*

1 Introduction

As if to achieve a primitive kind of sexual reproduction [56], retroviruses, including Human Immunodeficiency Virus (HIV), package two positive sense RNA molecules and a strand-switching reverse transcriptase enzyme in each virion. Even genetically distinct strains can be copackaged [27], and with documented cases of superinfection of both hosts [12, 20, 26] and cells [34], all ingredients necessary for recombination to mold the AIDS epidemic are in place [45]. 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 [59]. Viruses around the world have been classified into subtypes (e.g. A), circulating recombinant forms (e.g. CRF01_AE), and subsubtypes (e.g. A1) [50]. All major subtypes have long co-circulated in Africa, but subtype mixtures are increasingly common in other parts of the world (e.g. [6, 15]). With few barriers to limit recombination between genetically diverse strains [5, 11], 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 [9, 60]. HIV recombinants have successfully out-competed other strains within hosts [20, 52], and inter-subtype recombinants have been alarmingly successful at causing (e.g. [35, 46]) or taking over [55] local epidemics. Finally, early worries that recombination might facilitate the evolution of drug resistance have now been realized in treated patients [10, 43].

With recombination thus driving HIV evolution, it is important to uncover the mechanism of strand transfer. In-

1-4244-1509-8/07/$25.00 ©2007 IEEE

creasing experimental and observational evidence suggests that strand transfer events do not occur uniformly along the genome (see Fig. 1 for a map of the HIV genome). There have been several hints of a hotspot in the 5' portion of the *pol* gene, both in experiment [28] and among *in vivo* recombinants sampled from patients [36, 57]. Another well studied hotspot is the conserved C2 region of *env* [40, 48], although all conserved regions in *env* are relatively hot [4] and even variable regions can become hotspots with the right donor template [5]. Many inter-subtype recombinants observed around the world display a recombinant pattern where the variable loop V3, between C2 and C3 [49], or the complete gp120 portion of *env* [54] is swapped with another subtype. Other regions implicated as possible hotspots are 5' *gag* [18], the *gag-pol* boundary [36], the *pol-vif* boundary [16], through *vif* into the 5' *env* [36], a GC-rich region near the *tat/rev* splice site [17], and near *nef* [36]. 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 [23]). The recombination enhancing role of nucleocapsid [42] suggests the importance of secondary structure, and the TAR stem loop is one structure known to facilitate strand transfer [7, 29]. 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 [22, 21, 41]. Homology at the dimerization signal [11, 39] and throughout the genome enhances inter-molecular strand transfer [2], most importantly at the strand transfer site [5, 24]. It was recognized early that strand transfers tend to occur at pause sites of reverse transcription [61, 32], so it is not surprising that regions prone to form stem loops [62] and homopolymeric stretches [4] are positively correlated with strand transfer, since they are both associated with pause sites [30].

The study of *in vivo* recombination has revealed a nonuniform distribution of transfer sites along the genome [36, 49, 54, 57, 62]. 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 [4]. 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. Here, we have applied a hierarchical model for estimating spatial recombination rates [38] to a large dataset of recombinant sequences. While we report no new analytical methods, we discuss challenges encountered in the preparation and analysis of a large data set and present new results, including the identification of striking nonuniformities in the distribution of strand transfers along the HIV-1 genome.

2 Methods

2.1 Reference Alignment

We downloaded the 2003 reference sequences from the HIV database (www.hiv.lanl.gov), 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 [44], and gaps were trimmed from the alignment ends.

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 [19], which estimates, via Markov Chain Monte Carlo (MCMC), recombination in a single sequence [37]. We did not fix the tree relating potential parental subtypes in order to account for inconsistent relationships among subtypes along the genome (see, for example, Fig. 5 of [3]), which can bias recombination inference if neglected [19]. 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.

To overcome possible convergence issues resulting from these overly short MCMC runs, we prepared the remaining sequences for more thorough examination. 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 medians 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 [25] and all CODA [8] test statistics with default settings except in the Raftery & Lewis statistic, where convergence was diagnosed for the 0.025 quantile at a precision of ± 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.

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 [14] 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 [38], 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 [51] hyperprior on the parameters γ_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 [21], supports this correlation assumption.

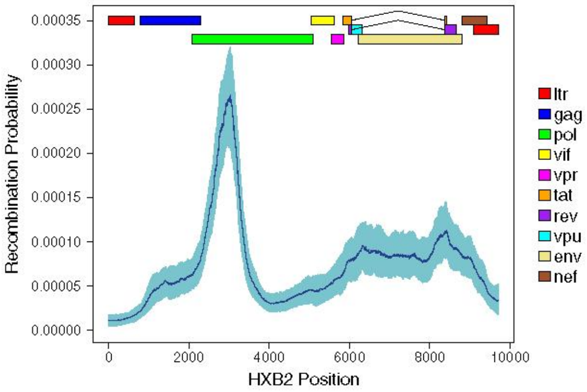

Figure 1. Spatial recombination profile of the full HIV-1 genome.

We prepared four data sets: (1) full genome, and three gene-focused datasets (2) *gag*, (3) *pol*, and (4) *env* to analyze with BandOfBrothers. 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 [31]. 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 [38].

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. For this analysis, we used the full genome data set. 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

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

where q_{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.

3 Results

The posterior distribution of recombination probabilities along the HIV-1 genome is summarized in Fig. 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. 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.

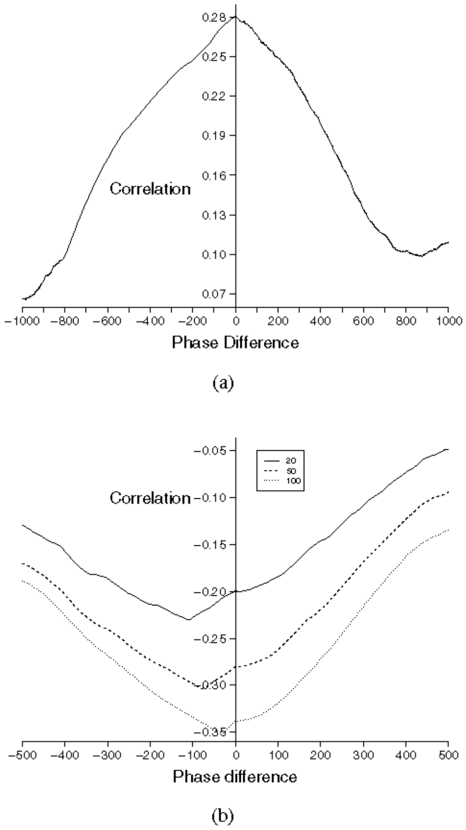

Figure 2. Lag correlation of recombination and (a) entropy and (b) GC content for window sizes 20,50, and 100.

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. 2a shows the correlation of recombination probability p_s with entropy H_s computed at each site s. Sequence variability is positively correlated with recombination both upstream and downstream of the strand transfer site. Fig. 2b shows correlation of GC content with recombination rate for different window sizes. There is a negative association of GC content and strand transfer, especially 50-100 nucleotides upstream of the transfer site. No correlations are large, especially considering the fact that the data points are not independent along the genome. Effective sample sizes estimated with CODA [8] 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. 3a) 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. 3b). Also shown in this plot are the locations of all *in vivo* drug resistance mutations mapping to *pol* [13]. 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 soley because these mutations represent the only variability in an otherwise conserved gene, we computed average pairwise similarity between dataset sequences. Similarity peaked just upstream of the recombination hotspot and was lowest in the 3' of *pol* (data not shown). Finally, the *env* analysis focused on the portion of *env* encoding gp120 (Fig. 3c) and revealed a concentrated recombination hotspot in the V3 loop.

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* [10, 43], 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 [1]. 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* [28, 36, 57], two [36, 57] 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 [28], although the viruses studied excluded the *env* gene. It is unlikely that widespread retroviral drug selection can fully explain the hotspot in *pol*, but whether the *pol* gene includes sequence with 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 [38]. Here, we analyzed 93 recombinants with all combinations of parents. Although different subtypes may vary in the preferred strand transfer sites [5], 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 [33]. Here, by clustering recombinants on their recombinant profile structure, we remove both recognized CRFs [33] 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 [40].

Recombination rates in the *env* gene were generally high (Fig. 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 [33, 49, 54]. Our gp120-focused results revealed a hotspot in V3, which is not visible in the full genome. 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 [53], 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 [5]. 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. 2). While correlations were not strong and probably not significant in our dataset, 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* [36]. We actually observed a

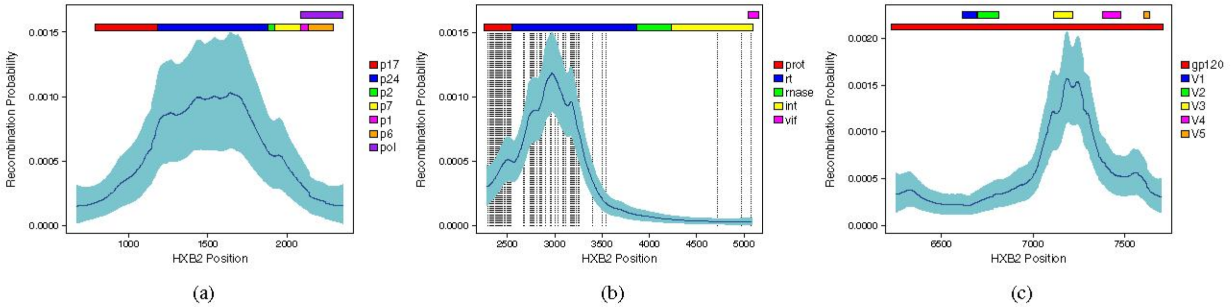

Figure 3. Spatial recombination profile for (a) *gag*, **(b)** *pol*, **and (c)** *env*

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 [5, 24]. 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 [47]. 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 [30]. 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 [58]. 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.

References

[1] *Towards Universal Access: Scaling up Priority HIV/AIDS Interventions in the Health Sector : Progress Report, April 2007.* World Health Organization, UNAIDS, UNICEF, 2007.

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

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

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

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

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

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

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

[9] D. S. Burke. Recombination in HIV: An important viral evolutionary strategy. *Emerg Infect Dis*, 3, 1997.

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

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

[12] B. Chohan, L. Lavreys, S. M. J. Rainwater, and J. Overbaugh. Evidence for frequent reinfection with Human Immunodeficiency Virus type 1 of a different subtype. *J Virol*, 79:10701–10708, 2005.

[13] S. A. Clark, C. Calef, and J. W. Mellors. *HIV Sequence Compendium*, chapter Mutations in Retroviral Genes Associated with Drug Resistance, pages 80–174. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 2005.

[14] M. Cornelissen, R. van den Burg, F. Zorgdrager, V. Lukashov, and J. Goudsmit. 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, 1997.

[15] M. T. Cuevas, I. Ruibal, M. L. Villahermosa, H. Daz, E. Delgado, E. V. de Parga, L. Prez-Alvarez, M. B. de Armas, L. Cuevas, L. Medrano, E. Noa, S. Osmanov, R. Njera, and M. M. Thomson. High HIV-1 genetic diversity in Cuba. *AIDS*, 16:1643–1653, 2002.

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

[17] N. W. Douglas, A. I. Knight, A. Hayhurst, W. Y. Barrett, M. J. Kevany, and R. S. Daniels. 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, 1996.

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

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

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

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

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

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

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

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

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

[27] W. S. Hu and H. M. Temin. 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, 1990.

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

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

[30] G. J. Klarmann, C. A. Schauber, and B. D. Preston. 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, 1993.

[31] B. Korber, B. T. Foley, C. Kuiken, S. K. Pillai, and J. G. Sodroski. *Human Retroviruses and AIDS Compendium*, chapter Numbering Positions in HIV Relative to HXB2CG, pages III–102–III–111. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1998.

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

[33] T. Leitner, B. Korber, M. Daniels, C. Calef, and B. Foley. *HIV Sequence Compendium*, chapter HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences, 2005, pages 41–48. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 2005.

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

[35] K. Liitsola, I. Tashkinova, T. Laukkanen, G. Korovina, T. Smolskaja, O. Momot, N. Mashkilleyson, S. Chaplinskas, H. Brummer-Korvenkontio, J. Vanhatalo, P. Leinikki, and M. O. Salminen. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS*, 12:1907–19, 1998.

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

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

[38] V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard. Phylogenetic mapping of recombination hot-spots in HIV

via spatially smoothed change-point processes. *Genetics*, 175:1773–1785, 2007.

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

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

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

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

[43] T. Nora, C. Charpentier, O. Tenaillon, C. Hoede, F. Clavel, and A. J. Hance. Contribution of recombination to the evolution of HIV expressing resistance to antiretroviral treatment. *J Virol*, 2007.

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

[45] M. Peeters. *HIV Sequence Compendium*, chapter Recombinant HIV Sequences: Their Role in the Global Epidemic, pages I39–I54. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 2000.

[46] S. Piyasirisilp, F. E. McCutchan, J. K. Carr, E. Sanders-Buell, W. Liu, J. Chen, R. Wagner, H. Wolf, Y. Shao, S. Lai, C. Beyrer, and X. F. Yu. 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, 2000.

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

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

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

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

[51] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

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

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

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

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

[56] H. M. Temin. Sex and recombination in retroviruses. *Trends Genet*, 7:71–74, 1991.

[57] M. M. Thomson, M. Sierra, A. Tanuri, S. May, G. Casado, N. Manjn, and R. Njera. 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, 2004.

[58] T. Toni, C. Adj-Tour, N. Vidal, A. Minga, C. Huet, M.-Y. Borger, P. Recordon-Pinson, B. Masquelier, M. Nolan, J. Nkengasong, H. J. Fleury, E. Delaporte, and M. Peeters. 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, 2005.

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

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

[61] W. Wu, B. M. Blumberg, P. J. Fay, and R. A. Bambara. 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, 1995.

[62] C. Y. Zhang, J. F. Wei, and H. S. H. 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, 2005.