

# Bayesian recombination identification: new models for incorporating prior information

Vladimir N. Minin<sup>1</sup>, Karin S. Dorman<sup>2</sup>, Marc A. Suchard<sup>1,3</sup>

<sup>1</sup> Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-1766, USA

<sup>2</sup> Departments of Statistics and Genetics, Development & Cell Biology and the Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011, USA

<sup>3</sup> Corresponding author: msuchard@ucla.edu

## Abstract

A dual multiple change-point (MCP) model was recently introduced to improve detection of recombination among aligned nucleotide sequences in a Bayesian framework. To improve the precision of recombination identification, one can employ informative priors on recombination locations. We develop a new class of prior distributions for the dual MCP model that flexibly incorporate information about previously identified recombinants. We suggest a simple rescaling procedure for our new prior to adjust the overall probability of recombination. Next, we demonstrate how to specify prior distributions for recombination locations using previous analyses of individual recombinants. Two possible types of prior information are considered: point-estimates and probabilities of recombination locations. In both cases we demonstrate how to overcome data sparseness by employing Gaussian Markov random fields. We examine two real examples of recombination analysis in HIV and apply our methodology to summarize spatial recombination information produced by these studies.

## 1 Introduction

Recombination is an important cause of the large genome rearrangements that punctuate the evolution of living organisms. Further, recombination remains paramount in rapidly evolving viruses that use it to adapt quickly to changing environmental conditions [Worobey and Holmes, 1999]. In human immunodeficiency virus (HIV), recombination impacts on several questions of medical importance, including drug resistance [Kellam and Larder, 1995] and vaccine development [Korber et al., 2001]. Recent findings suggest that recombination is a major force in the evolution of HIV with an *in vivo* rate of 1-2 recombination events per generation [Shriner et al., 2004]. The recombination frequency varies along the HIV genome suggesting the existence of regions where the process is promoted [Magiorkinis et al., 2003, Dykes et al., 2004, Zhuang et al., 2002]. Although the mechanisms generating such hot-spots are largely unknown, *in vivo* [Galletto et al., 2004] and *in vitro* [Moumen et al., 2003] studies support the involvement of RNA secondary structure. Strong interest in viral recombination has stimu-

lated both experimental [Dykes et al., 2004, Levy et al., 2004] and theoretical [Posada, 2002, Suchard et al., 2003, Husmeier and McGuire, 2003] research in the detection and mapping of recombination.

The non-uniform genomic distribution of recombination, together with emerging estimates of recombination locations, motivates the need for new models of recombination detection that are able to incorporate prior information from previous studies. A dual multiple change-point (MCP) model was recently introduced for accurate estimation of recombination locations [Minin et al., Submitted]. Here, we introduce a novel prior distribution over recombination locations that can be used to inject such prior information into the dual MCP model. We also give recommendations on how to rescale the distribution to control the overall prior probability of recombination. Summarizing prior information about recombination constitutes a challenge in its own right because there are relatively few known recombinant sequences. Point-estimates of recombination locations, usually reported as boundaries of the genomic regions with different parental heritage [Paraskevis et al., 2000, Levy et al., 2004], are

the most abundant source of prior information. Unfortunately, these estimates are sparsely distributed over the length of the HIV genome providing little information about the location of recombination hot-spots. Additionally, many studies reporting recombination locations fail to quantify the uncertainty in their estimates. Recent advances in phylogenetic recombination detection [Husmeier and McGuire, 2003, Suchard et al., 2002, Minin et al., Submitted] make calculating site-specific recombination probabilities possible. However, these methods still analyze one putative recombinant at a time, and new approaches are needed to average the results obtained from multiple recombinant sequences.

In this paper, we address the sparseness of available prior information by recruiting Gaussian Markov random fields (GMRFs). These well-studied models are successfully used in spatial statistics to take advantage of the highly structured, spatial data [Elliott et al., 2000]. Through a GMRF, we impose a biologically relevant correlation structure along the genome so that adjacent parts share information about recombination. Bayesian hierarchical models are adopted to accommodate two possible sources of information about recombination locations, point-estimates and site-specific probabilities of recombination. We use the results of earlier recombination studies to demonstrate the capabilities of the proposed models. Finally, in the Discussion, we examine issues that complicate interpretation of the results and outline future work that will help improve our understanding of HIV recombination.

## 2 Methods

### 2.1 Dual multiple change-point model

As with most phylogenetic methods of recombination detection [Hein, 1990, Husmeier and McGuire, 2003, Suchard et al., 2002], we start with a multiple sequence alignment of  $N$  homologous nucleotide sequences of length  $S$ , encoded by a matrix  $\mathbf{Y} = \{Y_{ns}\}$ , where  $n = 1, \dots, N$ ,  $s = 1, \dots, S$  and  $Y_{ns} \in \{A, G, C, T/U, -\}$ . Each site or column of the alignment,  $\mathbf{Y}_s = (Y_{1s}, \dots, Y_{Ns})^t$ , is assumed to evolve independently and is assigned a set of parameters  $\Phi_s$ , necessary to define a likelihood  $f(\mathbf{Y}_s | \Phi_s)$ . To specify the likelihood, we adopt a standard phylogenetic model that uses a continuous-time Markov chain over the state-space  $\{A, G, C, T\}$  to describe the substitution process resulting in the nucleotide polymorphism observed in each column [Felsenstein, 2004]. The infinitesimal transition rate matrix  $\Lambda_s$ , defining the Markov process, is parameterized in terms of its stationary distribution  $\pi_s = (\pi_{sA}, \pi_{sG}, \pi_{sC}, \pi_{sT})$  and a transition/transversion

rate  $\kappa_s \in [0, \infty]$  following Hasegawa et al. [1985]. To reduce the number of nuisance parameters in the model, we fix  $\pi_s$  to the observed nucleotide frequencies across the whole alignment. This leaves us with one free parameter  $\kappa_s$  defining the matrix  $\Lambda_s = \Lambda(\kappa_s)$ . To complete the phylogenetic model specification, we need a bifurcating tree topology  $\tau_s$ , describing the history of nucleotide substitutions, with branch lengths  $\mathbf{B}_s = (b_{1,s}, \dots, b_{2N-3,s})$  representing the expected number of substitutions between the bifurcation events. We further reduce the number of free parameters in the model by integrating  $\mathbf{B}_s$  out of the likelihood through assuming an exponential prior on each branch length  $p(b_{i,s}) \propto \exp(-b_{i,s}/\mu_s)$  for all  $i = 1, \dots, 2N - 3$ . In summary, the likelihood of site  $s$  is a function of three parameters  $\Phi_s = (\tau_s, \kappa_s, \mu_s)$ .

To characterize a recombinant, we need to infer the segments of the sequence that support different tree topologies when aligned against reference sequences. To accomplish this, we introduce a set of topology break-points  $1 = \theta_0 < \theta_1 < \dots < \theta_M < \theta_{M+1} = S + 1$  and fix  $\tau_s = \tau_m$ , for all  $s \in [\theta_{m-1}, \theta_m)$ . Topologies from adjacent segments are constrained to be distinct to assure that  $\theta$  represents a unique vector of recombination locations. The duality of this model comes into play when we introduce a similar set of change-points for substitution process parameters  $\mu_s$  and  $\kappa_s$ . Although substitution model parameters are not of direct interest during recombination detection, appropriate modeling their spatial variation improves the resilience of our model to false positive recombination identification [Dorman et al., 2002]. The specification of prior distributions for all parameters and details of Markov chain Monte Carlo (MCMC) sampling from the posterior distribution of the dual MCP model are described by Minin et al. [Submitted].

### 2.2 Informative prior for recombination locations

The dual MCP model is not only more accurate than previous models [Minin et al., Submitted], but also allows the construction of flexible site-specific prior probabilities of recombination

$$p_s = \Pr(s \in \{\theta_1, \dots, \theta_M\}) \text{ for } s = 2, \dots, S, \quad (1)$$

where  $p_s$  is the probability that site  $s$  is a recombination break-point. In the original dual MCP model all unconstrained topology break-points  $\theta_1, \dots, \theta_M$  are *a priori* uniformly distributed given  $M$ , while  $M$  itself follows a truncated Poisson prior distribution with mean  $\lambda$ . This formulation leads to a spatially uninformative prior with equal site-specific recombination probabilities

$$p_s = \frac{1}{S-1} \sum_{m=0}^{S-1} m \Pr(M = m) \approx \frac{\lambda}{S-1}, \quad (2)$$

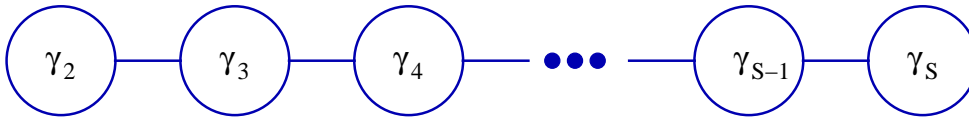


Figure 1: Graph describing the neighboring structure of the Gaussian Markov random field for  $\gamma = (\gamma_2, \dots, \gamma_S)$  along the genome.

where the above approximation holds since  $S \gg \lambda$ . We propose to directly estimate site-specific prior probabilities from existing spatial recombination information. Notice that in the site-specific model  $p_s$  control both the overall probability of recombination  $1 - \prod_{s=2}^S (1 - p_s)$  and its spatial preference. We show how to compute a scaling constant  $C$  to adjust the prior probability of recombination

$$1 - \prod_{s=2}^S (1 - Cp_s) = \delta, \quad (3)$$

for an arbitrary  $\delta \in (0, 1)$ .  $C$  can be adjusted to set  $\delta$ , often at 0.5 for an uninformative prior, while the transformed site-specific prior probabilities  $\tilde{p}_s = Cp_s$  retain the spatial information about recombination rates. We rely on the approximation  $\ln(1 - p) \approx -p$ , accurate for small  $p$ , to arrive at

$$\ln \prod_{s=2}^S (1 - \tilde{p}_s) = \sum_{s=2}^S \ln(1 - \tilde{p}_s) \approx -C \sum_{s=2}^S p_s. \quad (4)$$

Approximation (4) suggests that site-specific recombination probabilities, rescaled by  $C = -\ln(1 - \delta) / \sum_{s=2}^S p_s$ , approximately satisfy the desired condition (3).

### 2.3 Recombination location point-estimates

We now turn to the problem of using previous recombination studies to estimate prior site-specific probabilities  $p_s$ . Suppose that point-estimates of recombination break-point locations are available from  $K$  different recombinant sequences. We can map these location onto an alignment of the  $K$  sequences and associate with each site the total number of recombination occurrences  $Z_s$  and the total number of trials (opportunities for recombination)  $T_s \in \{1, \dots, K\}$ . This latter quantity represents the number of recombinant sequences that have homologous nucleotides at site  $s$  and varies with  $s$  due to insertions and deletions in the evolutionary process.

Assuming independence between recombination events,  $Z_s$  follows a binomial distribution

$$\Pr(Z_s = z | p_s, T_s) = \binom{T_s}{z} p_s^z (1 - p_s)^{T_s - z}. \quad (5)$$

With sufficient observations of recombination, we could directly estimate  $p_s$  as the frequencies of observed break-points across recombinants. However, in practice, the total number of observed recombination events is a few orders of magnitude less than the number of sites  $S$ . Therefore,  $\{Z_s\}$  constitutes a very sparse data set. To overcome this sparseness, we argue that site-specific recombination probabilities should have similar values at adjacent nucleotides. To model such a spatial dependency, we assume a GMRF prior on the site-specific recombination log-odds

$$\gamma_s = \log \left( \frac{p_s}{1 - p_s} \right). \quad (6)$$

GMRFs, also known as conditional auto-regressions, form a large class of models frequently used in image processing and spatial epidemiology [Besag et al., 1991, Knorr-Held and Rue, 2002]. They introduce spatial smoothing into a finite set of model parameters. If we specify an undirected graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with nodes  $\mathcal{V} = \{1, \dots, l\}$  and edges  $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V}\}$ , then random quantities at these nodes  $\mathbf{x} = (x_1, \dots, x_l)$  follow a GMRF with respect to  $\mathcal{G}$ , provided  $\mathbf{x} \sim N(\boldsymbol{\nu}, \mathbf{Q}^{-1})$  where  $\boldsymbol{\nu}$  is a mean vector and  $\mathbf{Q} = \{Q_{ij}\}$  is a precision matrix of a multivariate normal distribution, such that  $Q_{ij} \neq 0$  if and only if  $(i, j) \in \mathcal{E}$ .

Let  $\partial(i) = \{j : (i, j) \in \mathcal{E}\}$  denote the set of neighbors of the  $i$ -th node and let the notation  $\mathbf{x}_{\mathcal{A}} = \{x_i : i \in \mathcal{A}\}$  and  $\mathbf{x}_{-\mathcal{A}} = \{x_i : i \notin \mathcal{A}\}$ ,  $\mathcal{A} \subset \mathcal{V}$  refer to subsets of  $\mathbf{x}$ . Then, the non-zero pattern of the precision matrix  $\mathbf{Q}$  and the formula for the full conditional distribution of a single component

$$x_i | \mathbf{x}_{-i} \sim N(\nu_i - \sum_{j \neq i} \frac{Q_{ij}}{Q_{ii}} (x_j - \nu_j), Q_{ii}^{-1}) \quad (7)$$

imply the Markov property  $p(x_i | \mathbf{x}_{-i}) = p(x_i | \mathbf{x}_{\partial(i)})$ . Alternatively, one can start with the Markov properties to define a Markov field provided that all full conditional distributions are mutually compatible and a positivity condition is met [Besag, 1974, Besag and Kooperberg, 1995].

We use the intrinsic conditional autoregressive (ICAR) parameterization to model spatial interactions among recombination log-odds  $\gamma = (\gamma_2, \dots, \gamma_S)$  [Besag et al., 1991]. Figure 1 shows the graph representing spatial

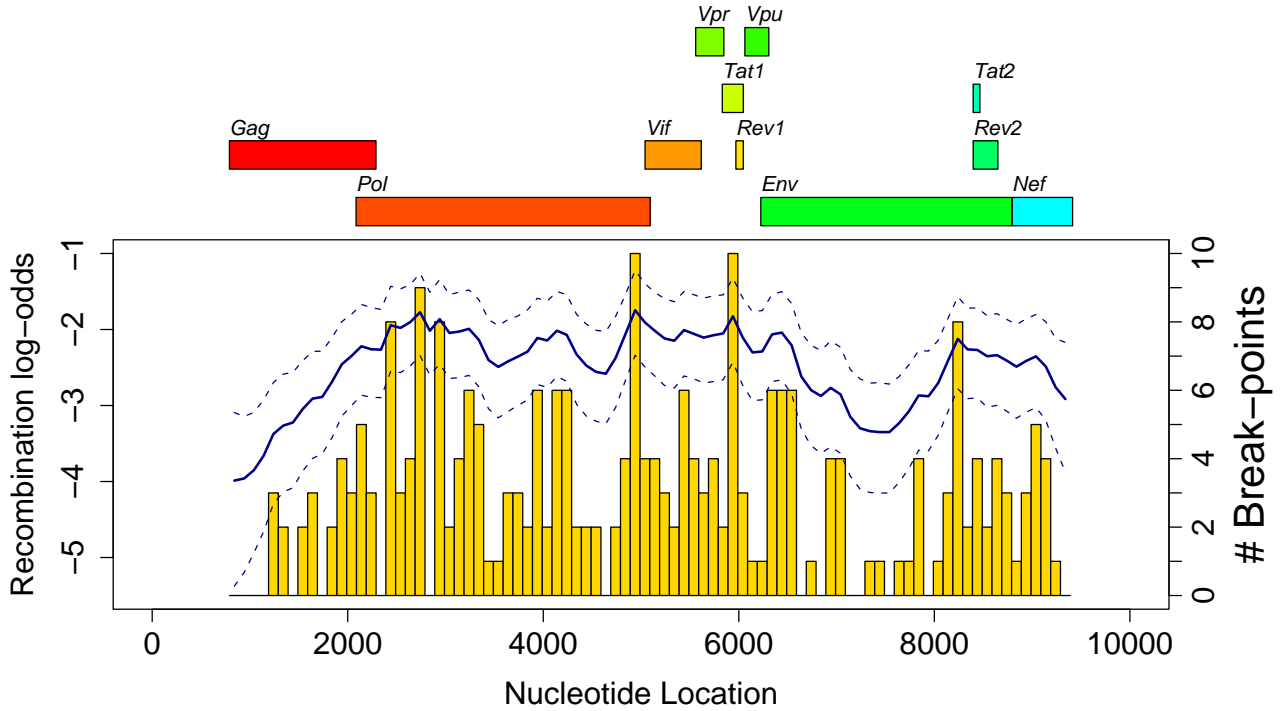


Figure 2: Recombination counts and inferred recombination log-odds. On the top the HIV gene map is depicted. The bottom plot shows counts of recombination occurrences (right vertical axis) as vertical yellow bars. Superimposed on them are posterior medians (solid line) and 95% confidence intervals (dashed lines) of recombination log-odds (left vertical axis).

dependencies of recombination locations. This linear graph is motivated by the linear processivity of reverse transcription during which recombination occurs. Our GMRF can be formulated by a set of full conditional distributions of components of  $\gamma$ :

$$\begin{aligned} \gamma_2 | \gamma_{-2} &\sim N(\gamma_3, \omega^{-1}), \\ \gamma_s | \gamma_{-s} &\sim N\left(\frac{\gamma_{s-1} + \gamma_{s+1}}{2}, \frac{1}{2\omega}\right) \\ &\text{for } s = 3, \dots, S-1, \text{ and} \\ \gamma_S | \gamma_{-S} &\sim N(\gamma_{S-1}, \omega^{-1}). \end{aligned}$$

A diffuse prior on the spatial precision parameter  $\omega \sim \Gamma(0.01, 0.01)$  completes our model specification. Expressing the distributions (8) in terms of a multivariate normal reveals a GMRF with mean  $\boldsymbol{\nu} = \mathbf{0}$  and precision matrix

$$\mathbf{Q} = \omega \times \begin{pmatrix} 1 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ 0 & & & & -1 & 1 \end{pmatrix}. \quad (8)$$

In the view of equality  $\mathbf{Q}\mathbf{1} = \mathbf{0}$ , it is clear that matrix  $\mathbf{Q}$  is not positive definite, leading to an improper joint prior distribution for  $\gamma$ . However, the propriety of the posterior distribution has been shown for generalized linear

models under mild conditions that are usually satisfied in the presence of any informative data [Ghosh et al., 1998, Sun et al., 1999]. Although other GMRF formulations are possible, the ICAR model is the most popular choice when imposing strong spatial correlations [Besag et al., 1991]. The posterior distribution of all model parameters

$$\begin{aligned} &\Pr(\boldsymbol{\gamma}, \omega | \mathbf{Z}) \\ &\propto \Pr(\mathbf{Z} | \boldsymbol{\gamma}) \Pr(\boldsymbol{\gamma} | \omega) \Pr(\omega) \\ &\propto \prod_{s=2}^S \left(\frac{e^{\gamma_s}}{1 + e^{\gamma_s}}\right)^{Z_s} \left(\frac{1}{1 + e^{\gamma_s}}\right)^{T_s - Z_s} \times \quad (9) \\ &\omega^{(S-2)/2} \exp\left(-\frac{\omega}{2} \sum_{s=2}^{S-1} (\gamma_s - \gamma_{s+1})^2\right) \times \\ &\omega^{0.01-1} e^{-\omega \times 0.01}, \end{aligned}$$

can be derived only up to a proportionality constant. As a consequence, we employ MCMC methodology to generate a random sample from distribution (9) and then draw inference from summaries of this finite sample.

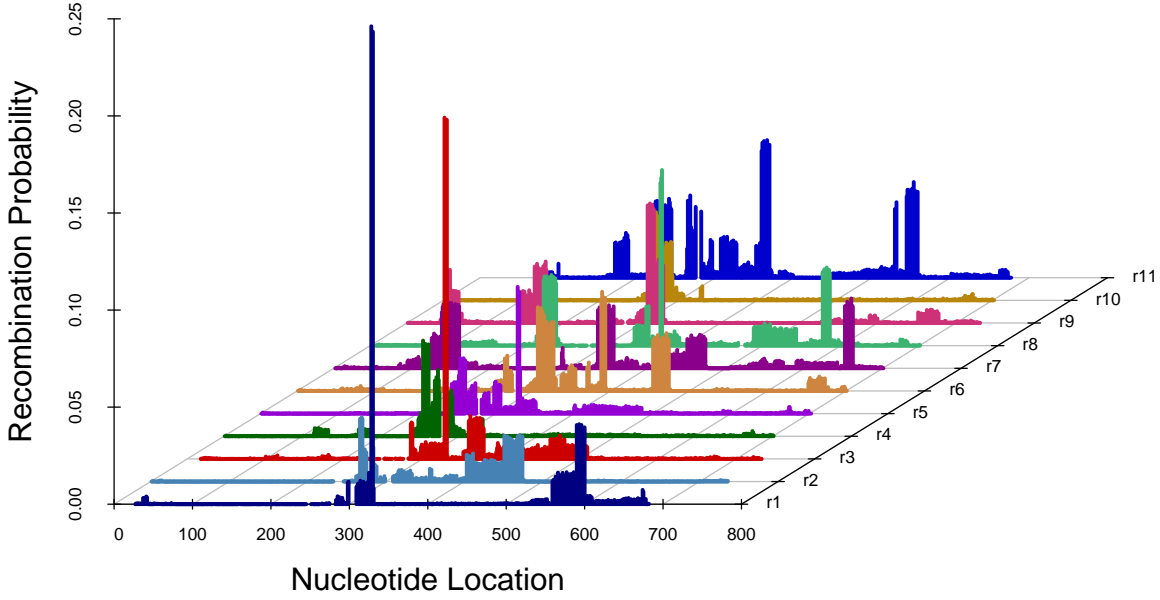


Figure 3: Results of individual analyses of 11 *gag* recombinants. Recombination probabilities are plotted for each site along the gene and for each recombinant. Recombinants are assigned distinct colors.

## 2.4 Site-specific recombination probabilities

Suppose that we have an alignment of length  $S$  with  $K$  recombinant sequences, but instead of recombination locations, a table  $\mathbf{P} = \{p_{ks}\}$ ,  $k = 1, \dots, K$ ,  $s = 2, \dots, S$  of site-specific recombination probabilities is provided. Notice that gaps in the alignment result in missing entries in table  $\mathbf{P}$ . We do not impute the missing data in our calculations, since if a recombinant lacks a particular nucleotide in its genome, recombination could not have occurred at that site. To simplify notation, we rearrange table  $\mathbf{P}$  so that each column contains only non-missing values allowing for unequal number of observations across columns. We assume that observed log-odds of recombination at each site are normally distributed about an unknown site-specific mean  $\eta_s$  with variance  $\psi^{-1}$ ,

$$\log\left(\frac{p_{ks}}{1-p_{ks}}\right) \sim N(\eta_s, \psi^{-1}). \quad (10)$$

Although this model postulates a common variance across sites, heterogeneity in recombination probabilities is achieved by allowing the means  $\boldsymbol{\eta} = (\eta_2, \dots, \eta_S)$  to differ. To address the sparseness of the data, we assign an ICAR prior for site-specific means  $\boldsymbol{\eta}$  as in (8). After specifying a diffuse prior for  $\psi \sim \Gamma(0.01, 0.01)$ , the posterior distribution becomes

$$\Pr(\boldsymbol{\eta}, \psi, \omega | \mathbf{P}) \propto \Pr(\mathbf{P} | \boldsymbol{\eta}, \psi) \Pr(\boldsymbol{\eta} | \omega) \Pr(\omega) \Pr(\psi). \quad (11)$$

The main difference between posterior distributions (11) and (9) is the definition of the likelihood

$$\prod_{s=2}^S \psi^{T_s/2} \exp\left\{\sum_{k=1}^{T_s} \left[\log\left(\frac{p_{ks}}{1-p_{ks}}\right) - \eta_s\right]^2\right\}, \quad (12)$$

where  $T_s$  is the number of non-missing entries in column  $s$ . Notice, that the Gaussian form of (12) and the ICAR prior on  $\boldsymbol{\eta}$  imply that full conditional distribution of  $\boldsymbol{\eta}$  is also multivariate normal. Hence, it is possible to implement a Gibbs sampler for all model parameters, leading to more efficient exploration of the posterior distribution (11). For both models, we use WinBugs [Spiegelhalter et al., 1999] for posterior simulation.

## 3 Results

We first apply our methodology to the recombination point-estimates reported by Magiorkinis et al. [2003]. These authors examined 35 full-length HIV genomes of putative recombinants and identified 247 break-points by bootscanning [Lole et al., 1999]. Poor spatial resolution of recombination detection with bootscanning does not allow us to consider site-specific recombination estimates. Instead we group nucleotides into 86 consecutive blocks, each containing 100 sites. We record the number of recombinations per block  $Z_s$  for  $s = 1, \dots, 86$  and use them as our dataset to infer recombination log-odds  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_S)$ . A histogram of these counts is depicted in Figure (2) in yellow. Additionally the posterior distribution of recombination log-odds  $\boldsymbol{\eta}$  is summarized

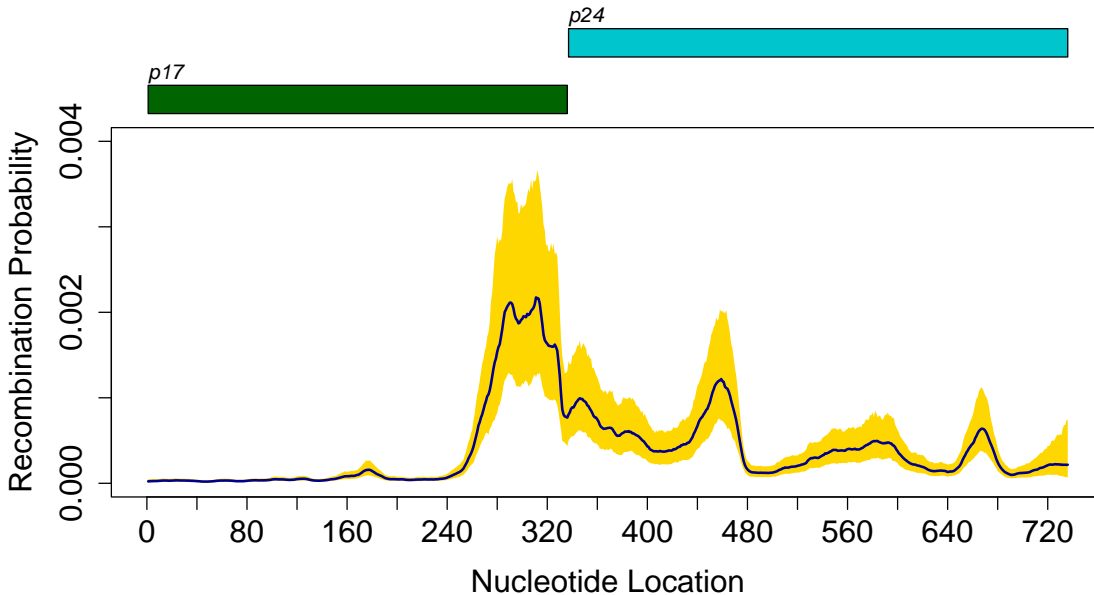


Figure 4: Probabilities of recombination averaged over individual *gag* sequences estimated in the presence of an ICAR prior. The locations of gene products within the *gag* gene are plotted on the top. The bottom plot contains posterior medians (solid blue) and 95% Bayesian credible intervals (yellow shading) of recombination log-odds  $\eta_s$ ,  $s = 2, \dots, S$ , transformed back to probability space.

with the posterior medians (solid blue) and 95% Bayesian credible intervals (dashed blue). This data set does not contain sufficient information to draw conclusions about hot-spot locations. Nevertheless, it appears that a region in the middle of the *env* gene is less likely to recombine than the rest of the genome. This domain of low recombination occurrence overlaps with the surface glycoprotein *gp120* gene whose genetic hyper variability HIV exploits to evade the host immune response. This finding is consistent with the positive correlation of recombination locations and sequence similarity reported by Magiorikinis et al. [2003].

In our second example, we concentrate on recombination within the *gag* gene of HIV. This gene spans about 700 nucleotides. Yirrell et al. [2002] collected 11 *gag* recombinant sequences in Uganda. We independently analyzed each sequence with the dual MCP model to arrive at site-specific probabilities of recombination and collected these estimates into matrix  $\mathbf{P}$ . We plot the entries of  $\mathbf{P}$  in Figure (3) as vertical bars. The horizontal plane of this three dimensional plot contains a two dimensional grid of nucleotide indices and recombinant labels. Colors in Figure (3) are used to help visualize different recombinant sequences. Next, we apply our model for site-specific recombination probabilities to the results of the dual MCP analyses in order to summarize the pattern of recombination occurrences in the *gag* gene. Figure (4) shows protein products *p17* and *p24* within the *gag* gene at the top and results of our analysis in the

bottom plot. In this plot, we show posterior medians (solid blue) and 95% Bayesian credible intervals (yellow shading) of recombination log-odds  $\eta_s$ , transformed back to probability space,  $\bar{p}_s = e^{\eta_s} / (1 + e^{\eta_s})$ . There is an apparent region of high recombination probability near the boundary between *p17* and *p24*.

## 4 Discussion

We introduce a new class of priors for recombination locations along a genome that can be used with the dual MCP model [Minin et al., Submitted]. After suggesting a simple rescaling procedure for our prior, we propose two models to summarize previous recombination detection studies to specify this prior. In both models, we recruit GMRFs to overcome the sparseness of existing recombination estimates. Motivated by Besag’s interpretation of the ICAR parameterization as a “stochastic version of linear interpolation” [Besag et al., 1991], we use the GMRF to interpolate recombination log-odds at sites where recombination has not been previously documented. The spatial correlation structure that we impose on recombination locations has a meaningful biological interpretation. In retroviruses, recombination occurs during template switching by reverse transcriptase as it linearly copies the viral RNA genome into DNA [Negroni and Buc, 2001]. Therefore, we expect adjacent nucleotides to have similar probabilities of recombination.

However, base pairing between distant nucleotides in the RNA secondary structure may also introduce more complicated spatial patterns of recombination probabilities. These long range interactions can be incorporated into our GMRF model by adding appropriate edges to the graph describing the neighboring structure of the Markov field.

Our results of pooling information from several recombinants should be interpreted with caution. If the analyzed sequences originated from independent recombination events and one neglects selection forces, regions with estimated high probability of recombination may be interpreted as hot-spots. However, it is possible that several seemingly different recombinant sequences share a common ancestry and are the product of a single recombination event more distant in the past. It is extremely difficult to distinguish between these two possibilities. It is common to conclude that sequences are descendents of the same recombination event and discard all of them but one, if the distance between their break-points is small [Fang et al., Submitted]. In the view of the possible existence recombination of hot-spots, this approach may be too conservative. New statistical tools are needed to test these hypotheses.

Finally, to more efficiently use the information in a moderate number of sequences, we plan to develop a Bayesian hierarchical approach that simultaneously estimates recombination break-points and their prior distribution. Such a strategy pools information not only along the genome, but also across different recombinants. A further advantage of the hierarchical GMRF prior is that additional covariates could be included into the model. We plan to test if several local sequence features are associated with recombination. We hope that using the right covariates will help estimate break-points more efficiently and shed light on the biological mechanisms leading to non-uniform distribution of recombination events along the genome.

## 5 Acknowledgements

We would like to thank Dimitris Paraskevis for providing us with estimates of recombination break-points from full-length HIV genomes. This work was supported by NIH grant GM068955.

## References

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

- J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746, 1995.
- J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.
- K.S. Dorman, A.H. Kaplan, and J.S. Sinsheimer. Bootstrap confidence levels for HIV-1 recombinants. *Journal of Molecular Evolution*, 54:200–209, 2002.
- 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.
- P. Elliott, J.G. Wakefield, N.G. Best, and D.J. Briggs, editors. *Spatial Epidemiology: methods and applications*. Oxford University Press, 2000.
- F. Fang, M.A. Suchard M. Rischmiller, and K.S. Dorman. Recombination and crossover point identification in full-length hepatitis B viruses. *Journal of Virology*, Submitted.
- J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA, 2004.
- R. Galetto, A. Moumen, V. Giacomoni, M. Veron, and P. Charneau amd M. Negroni. The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot *in vivo*. *Journal of Biological Chemistry*, 279:36625–36632, 2004.
- M. Ghosh, K. Natarajan, T.W.F. Stroud, and B.P. Carlin. Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93:273–282, 1998.
- M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.
- J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98:185–200, 1990.
- D. Husmeier and G. McGuire. Detecting recombination in 4-taxa DNA sequence alignment with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, 20:315–337, 2003.
- P. Kellam and B.A. Larder. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *Journal of Virology*, 69:669–674, 1995.

- L. Knorr-Held and H. Rue. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29:597–614, 2002.
- B. Korber, B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours. Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin*, 58:19–42, 2001.
- D.N. Levy, G.M. Aldrovandi, O. Kutsch, and G.M. Shaw. Dynamics of HIV-1 recombination in its natural target cells. *Proceedings of the National Academy of Sciences, USA*, 101:4204–4209, 2004.
- K.S. Lole, R.C. Bollinger, R.S. Paranjape, D. Gadkari S.S. Kulkarni, N.G. Novak, R. Ingersoll, H.W. Sheppard, and S.C. Ray. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of Virology*, 73:152–160, 1999.
- G. Magiorkinis, D. Paraskevis, A. Vandamme, E. Magiorkinis, V. Sypsa, and A. Hatzakis. *In vivo* characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. *Journal of General Virology*, 84:2715–2722, 2003.
- V.N. Minin, K.S. Dorman, and M.A. Suchard. Decoupling of change-points leads to more accurate recombination detection. *Bioinformatics*, Submitted.
- A. Moumen, L. Polomack, T. Unge, M. Veron, H. Buc, and M. Negroni. Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA. *Journal of Biological Chemistry*, 278:15973–15978, 2003.
- M. Negroni and H. Buc. Mechanisms of retroviral recombination. *Annual Review of Genetics*, 35:275–302, 2001.
- D. Paraskevis, M. Magiorkinis, V. Papanizos, G.N. Pavlakis, and A. Hatzakis. Molecular characterization of a recombinant HIV type 1 isolate (A/G/E/?): unidentified regions may be derived from parental subtype E sequences. *AIDS Research and Human Retroviruses*, 16:845–855, 2000.
- D. Posada. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular Biology and Evolution*, 19:708–717, 2002.
- D. Shriner, A.G. Rodrigo, D.C. Nickle, and J.I. Mullin. Pervasive genomic recombination of HIV-1 *in vivo*. *Genetics*, 167:1573–1583, 2004.
- D.J. Spiegelhalter, A. Thomas, and N.G. Best. *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit, 1999.
- M.A. Suchard, R.E. Weiss, K.S. Dorman, and J.S. Sinsheimer. Oh brother, where art thou? a Bayes factor test for recombination with uncertain heritage. *Systematic Biology*, 51:715–728, 2002.
- M.A. Suchard, R.E. Weiss, K.S. Dorman, and J.S. Sinsheimer. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. *Journal of the American Statistical Association*, 98:427–437, 2003.
- D. Sun, R.K. Tsutakawa, and P.L. Speckman. Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika*, 86:341–350, 1999.
- M. Worobey and E.C. Holmes. Evolutionary aspects of recombination in RNA viruses. *Journal of General Virology*, 80:2535–2543, 1999.
- D.L. Yirrell, P. Kaleebu, D. Morgan, C. Watera, B. Magambo F. Lyagoba, and J. Whitworth. Inter- and intra-genic intersubtype HIV-1 recombination in rural and semi-urban Uganda. *AIDS*, 16:279–286, 2002.
- J. Zhuang, A.E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B.D. Preston, and J.P. Dougherty. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of Virology*, 76:11273–11282, 2002.