Advanced Bayesian Phylogenetics: Recombination

Philippe Lemey and Marc A. Suchard

Rega Institute

Department of Microbiology and Immunology

K.U. Leuven, Belgium, and

Departments of Biomathematics and Human Genetics

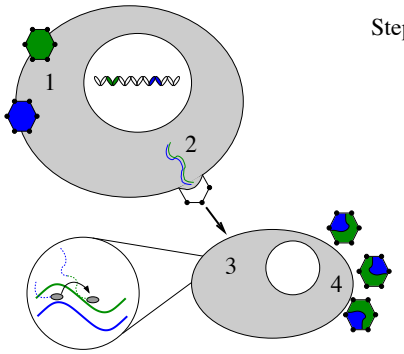
David Geffen School of Medicine at UCLA

Department of Biostatistics

UCLA School of Public Health

Genomic Reassortment in HIV

Dual infection can lead to inter-subtype recombination:



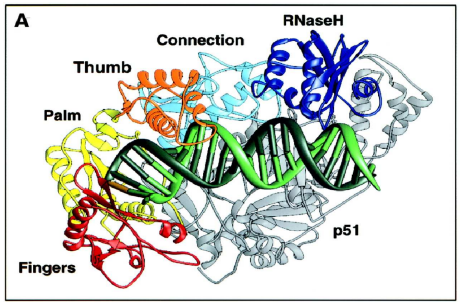
Steps in HIV Sex:

1. Co–infection of host cell by two different HIV subtypes
2. Co–packing of two different HIV RNAs into a single viron
3. Strand jumping during reverse transcription in newly infected host
4. Release of recombinant virus

- Originally believed **rare**
- Now suspected as **major** contributor to genetic diversity
- Clinical implications of intra-host recombination

Mechanisms and Hot-Spots?

Reverse Transcriptase



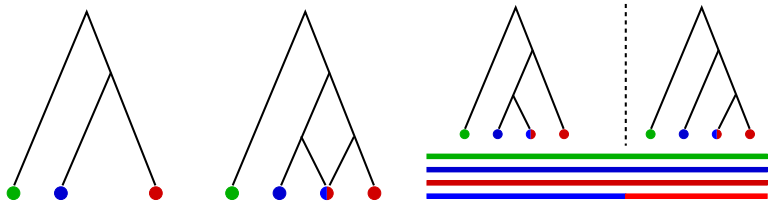
- Probability of strand-slippage/jumping may be function of genomic 1° or 2° structure

Clinical Relevance of Hot-Spots

- *In vivo* evidence
- *In vitro* recomb. rates $50\times$ greater in *env* than *gag*
- Development of multiple drug resistance (Kitchen *et al.*, 2006)
- Drug choice

Phylogenetic Recombination Detection

Recombination $\Rightarrow$ genomic break-points with **incongruent** topologies:

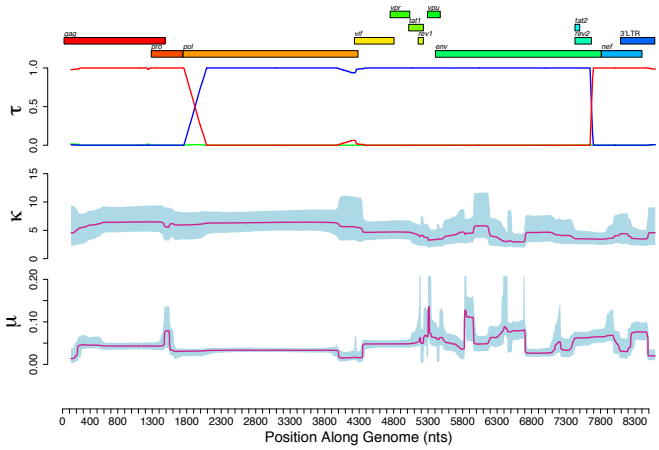


Map break-point locations:

- Hidden Markov models (Husmeier *et al.*, 2003, 2005)
- Multiple change-point models (Suchard, Minin *et al.*, 2002, 2003, 2005)

Dual Multiple Change-Point Process

KAL153
(AB recombinant)

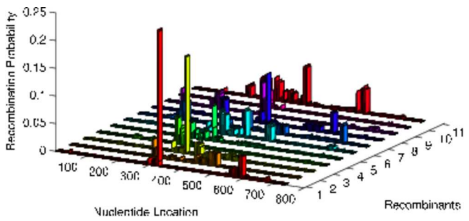


- Separate break-points and rate change-points
- Uncertainty on **number** and **locations** of break-points
- Variable dimensional model $\Rightarrow$ reversible jump MCMC

Putative A/G Inter-Subtype Recombinants

Data:

- 42 **unrelated** (hopefully) recombinants from LANL
- Of **African** origins
- Same subtypes to maximize power



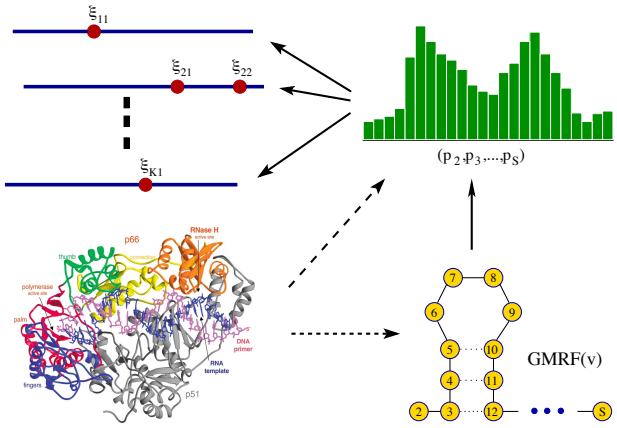
Subset of **independent** analyses

How to **pool** information?

- Sparse "observations" (# break-points $<<$ seq. length)
- Neighboring sites should have similar probabilities

Joint Analysis via Gaussian Markov Random Fields

A GMRF to smooth and estimate **population-level** recombination log-odds (probabilities):



Normally distributed vector

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\nu}, \mathbf{Q}^{-1})$$

is a GMRF wrt graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ when $\mathbf{Q} > 0$ and $Q_{ij} \neq 0$ iff $(i, j) \in \mathcal{E}$

- $\mathbf{Q}$ can be huge, but very sparse
- Fast numerical methods available, make the approach feasible (Rue *et al.*, 2001, 2004)

GMRF as an Improper Prior

Field (population-level log-odds) $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}|\omega \sim \mathcal{N}(\prime, \tilde{\mathbf{Q}}^{-1}), \text{ where } \tilde{\mathbf{Q}}^{-1} = \mathbf{Q} + \epsilon\mathbf{I}$$

Impropriety: 1^{st}-order random-walk field defined on **differences**. Baseline $\propto 1$. Normally, not a **problem** (Sun, 1999).

- Think of break-points as "success counts" $\mathbf{C} = (C_1, \ldots, C_S)$ in binomial trials
- What if $C_s = 0$ or $C_s = 42$ for all s?

Prior: Random-walk precision $\omega \sim \Gamma(\cdot, \cdot)$. Express prior belief via $p_i/p_j \leq 7$-fold (Bernardinelli, 1995; Moumen, 2001)

Non-linearly Constrained GMRFs

The number of break-points $M \sim$ approximately Poisson(δ) with $\delta = \sum_{s=1}^{S} p_s$ (le Cam, 1960) for each recombinant.

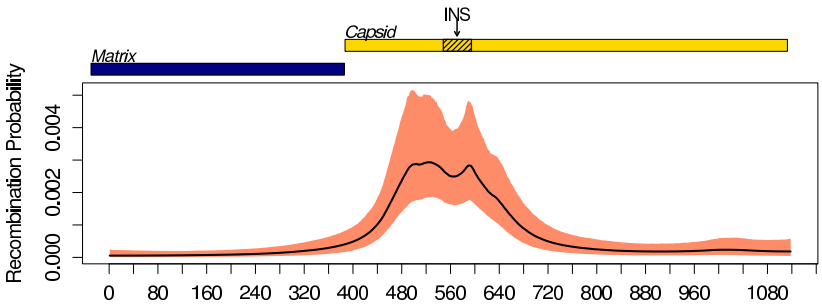
Aim: $\Pr(M > 0) \approx 1 - e^{\delta} = c = 0.5$.

Problem: Sum-of-$\mathbf{p}$ constraint is **non-linear** in the field (γ):

$$\sum_{s=1}^{S} \frac{e^{\gamma_s}}{1 + e^{\gamma_s}} = -\ln(1 - c) \tag{1}$$

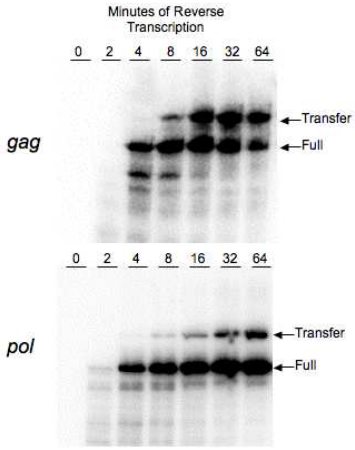
Solution: Linearize constraint via Taylor expansion about arbitrary point $\mathbf{v}$, then constraint $\Rightarrow$ "re-centering" proposal $\gamma^{\star}$ from unconstrained GMRF. How to choose $\mathbf{v}$?

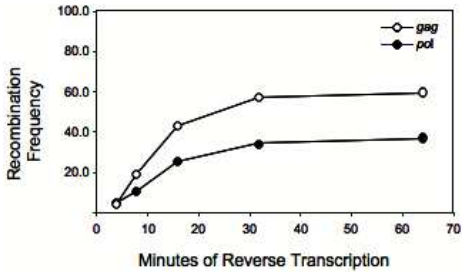
Hot-Spot in *gag* Gene



- Spatial association with an **instability element** (INS). INSs are motifs involved in post-transcriptional regulation of gene expression
- *In vivo* confirmation of hot-spot in the works

Preliminary *in vitro* Strand Transfer Assay





- Results **support** *gag* hot-spot. First *de novo* elucidation of HIV recombination mechanism with computational methods?