Conference on the Evolutionary Genetics of Infectious Disease

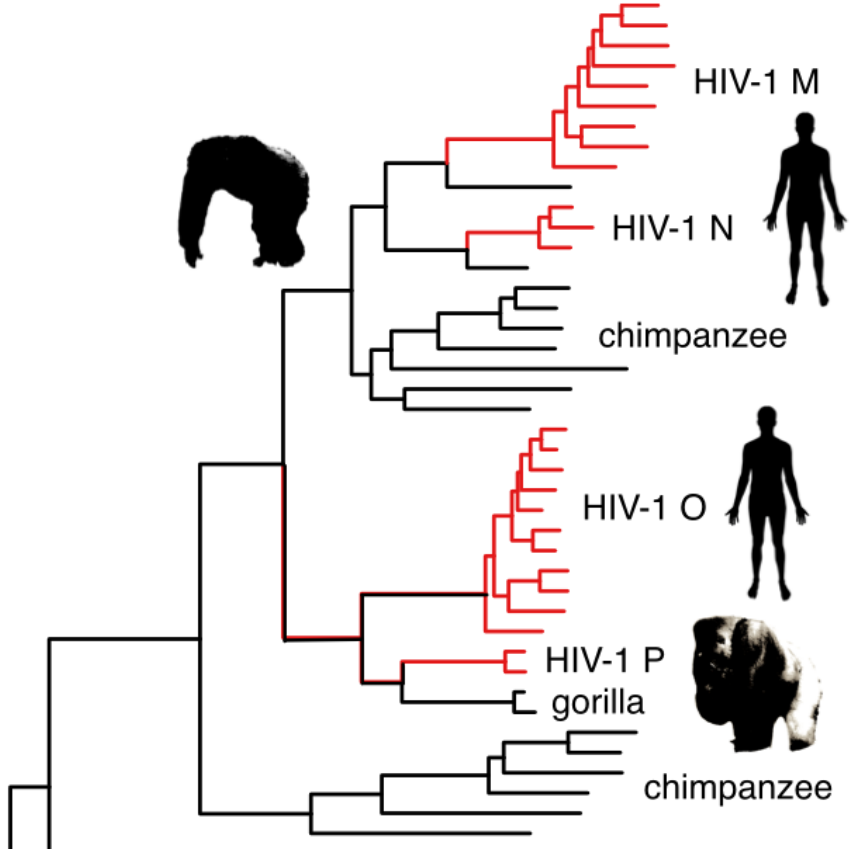
Unsupervised excursions into the deep evolutionary history of HIV-1 group M

Art Poon

Western University

Department of Pathology and Laboratory Medicine, Department of Applied Mathematics, Department of Microbiology and Immunology

Phylogeny of HIV/SIV *gag*



Excerpt of figure from Joy *et al* (2015) Global Virology I, Springer. Generated by maximum likelihood.

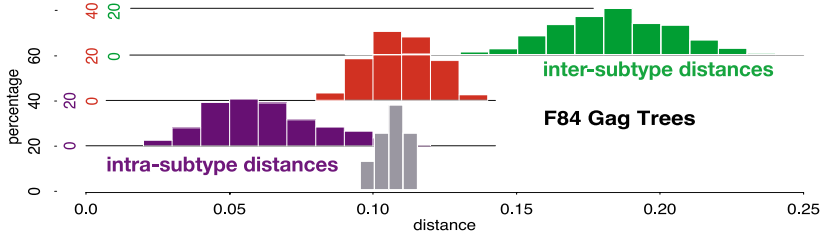
HIV-1 subtypes

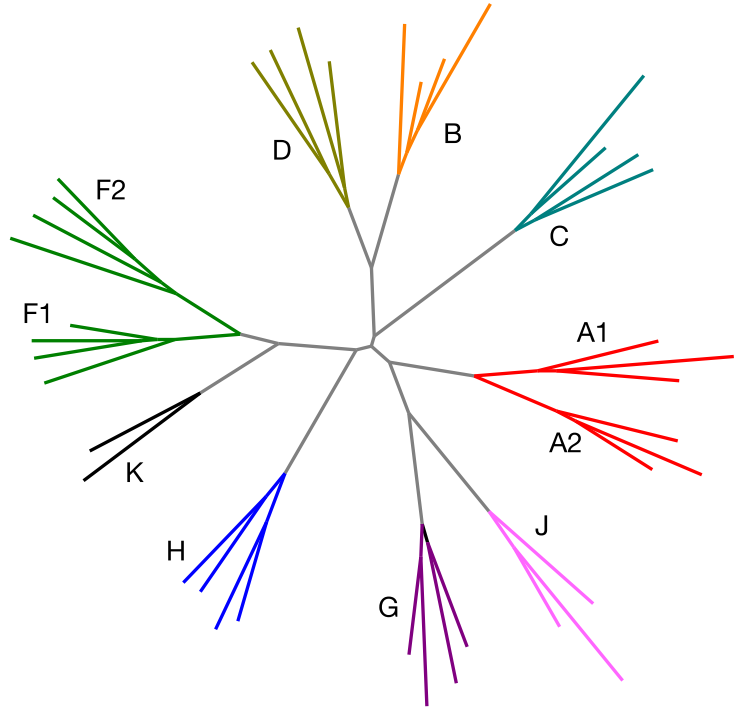
- Extraordinary diversity structured into subtypes.

HIV-1 Nomenclature Proposal

A Reference Guide to HIV-1 Classification

D.L. Robertson[1], J.P. Anderson[2], J.A. Bradac[3], J.K. Carr[4], B. Foley[5], R.K. Funkhouser[6], F. Gao[7], B.H. Hahn[7], C. Kuiken[5], G.H. Learn[2], T. Leitner[8], F. McCutchan[4], S. Osmanov[9], M. Peeters[10], D. Pieniazek[11], M.L. Kalish[11], M. Salminen[12], P. Sharp[13], S. Wolinsky[14], and B. Korber[5,6]





(*above*) ML tree from LANL curated (2010) HIV-1/M env subtype references, aligned with MAFFT (excluding indel-rich regions) and reconstructed with RAxML.

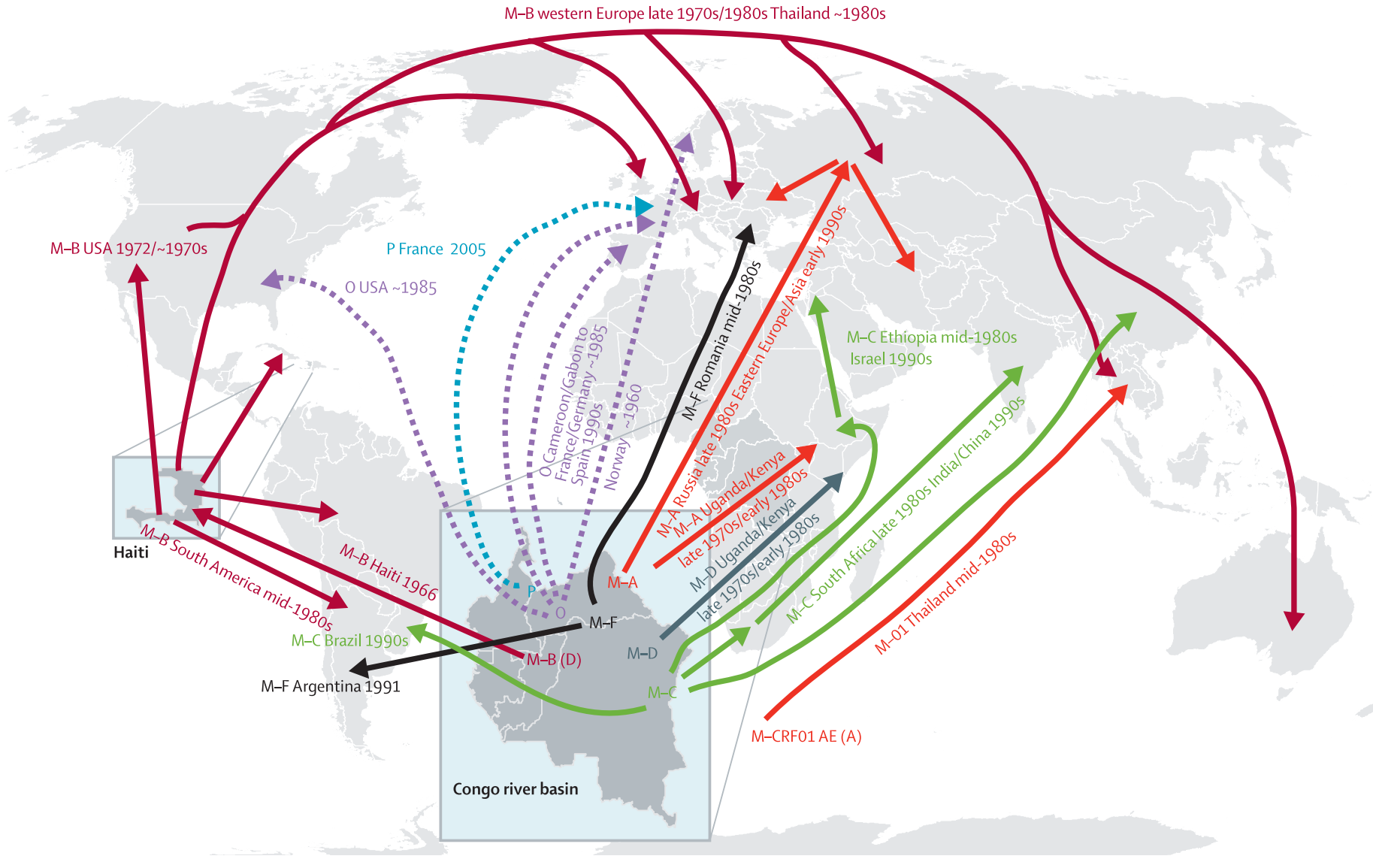
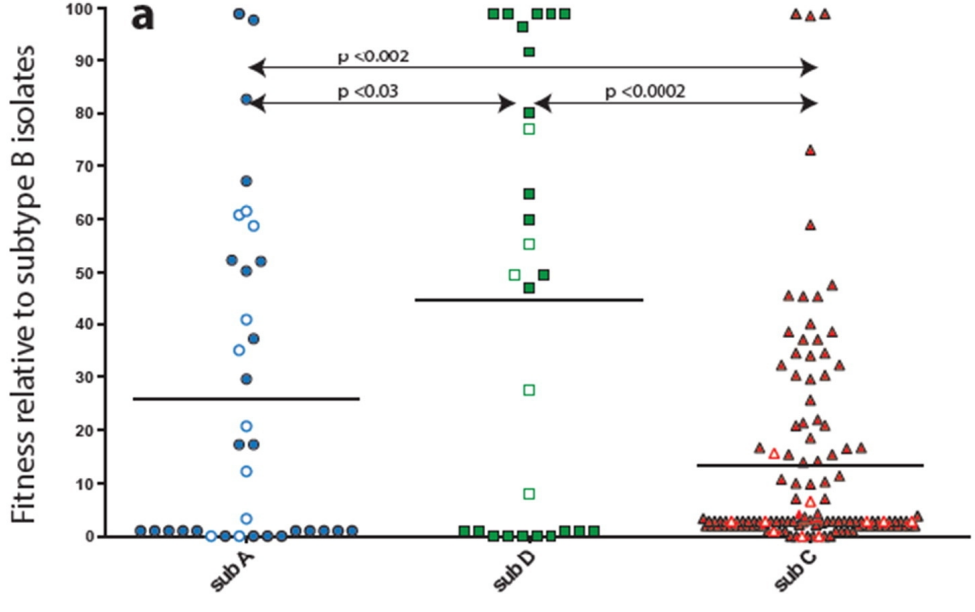


Figure from Tebit and Arts (2011) Lancet Inf Dis 11: 45-56.

Impact of HIV-1 subtypes

- Understanding the global epidemiology of HIV-1/M
- Variation in rates of progression to AIDS (subtypes C, D)



- Distinct mutational pathways to evolve drug resistance.

Figure from C Venner *et al.* (2016) EBioMedicine 13: 305-314.

Subtyping

- Compare a sequence to a reference database of "pure" subtype sequences
- Identify recombinant as a combination of two or more parent reference sequences.
- Sliding window methods (BOOTSCAN, REGA)
- Phylogenetic methods (SCUEAL)
- Linear classifiers (COMET)

Recombinant history of SIV

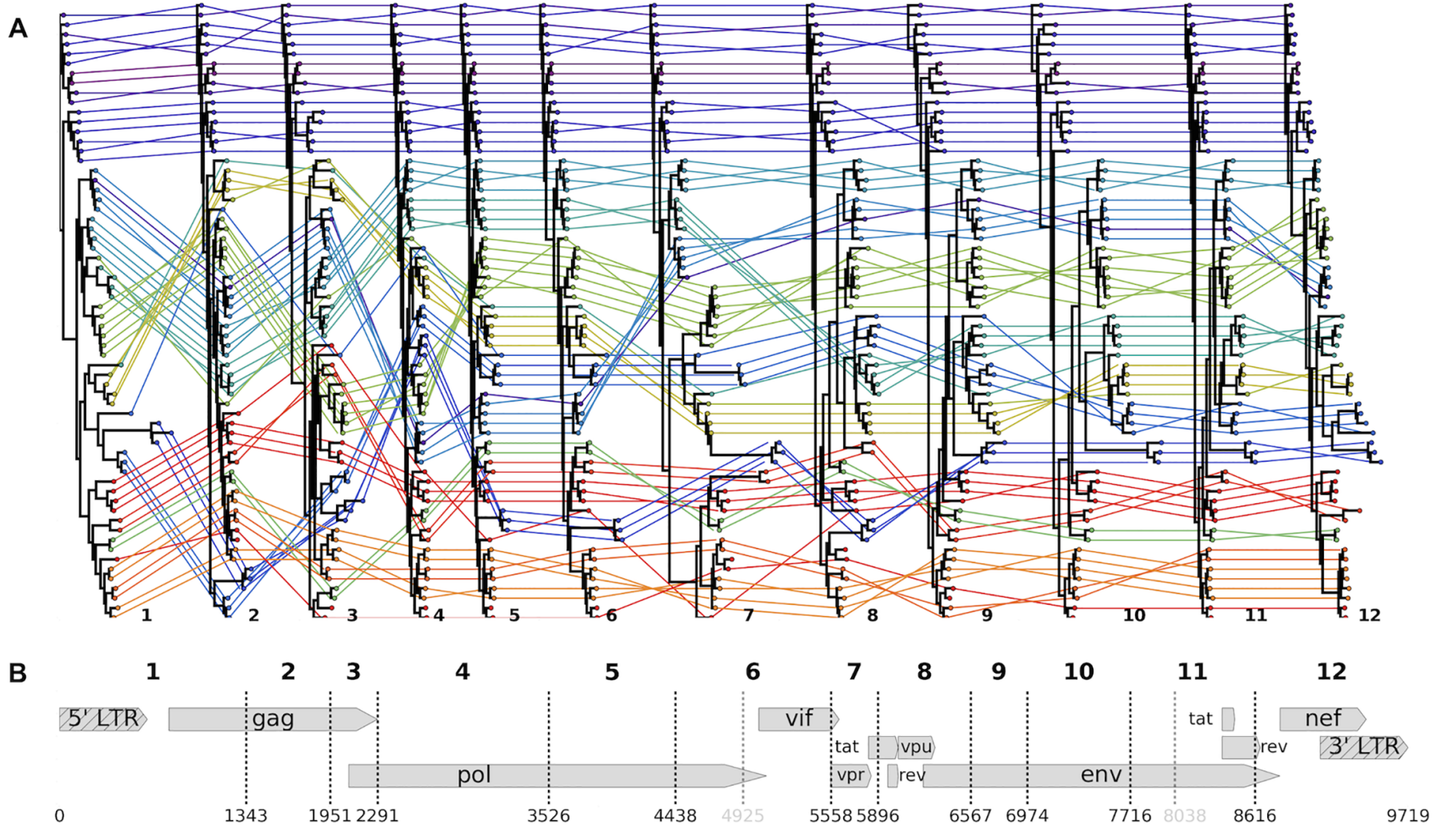


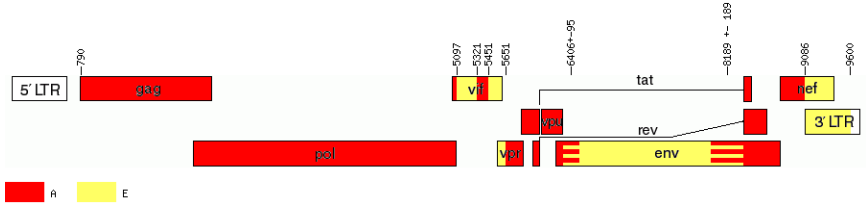
Figure from Bell and Bedford (2017) PLOS Pathog 13: e1006466.

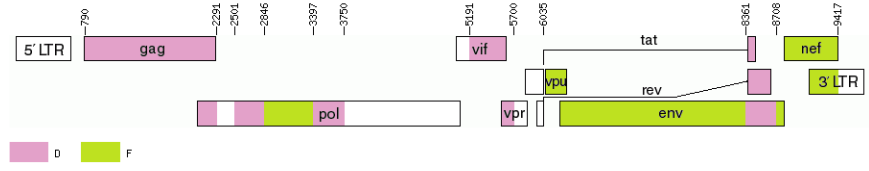
HIV-1 recombination

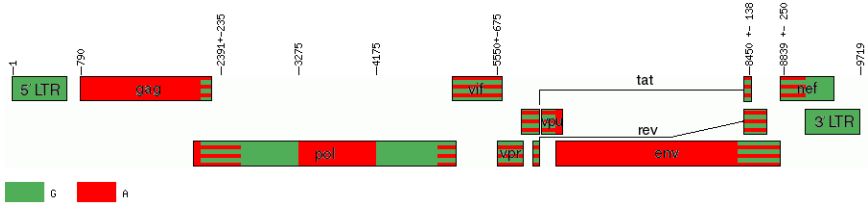
- Virus carries two copies of its genome.
- HIV reverse transcriptase switches templates during infection cycle.
- If a host cell is multiply infected, progeny virus can be recombinant.
- Effective recombination rate estimated at about 1%/genome/generation[1].
- Close to estimated mutation rate (0.2/genome/generation)[2].

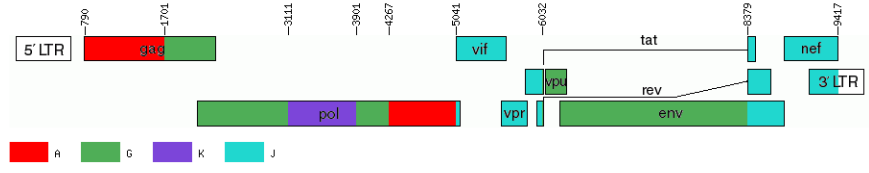
[1] RA Neher *et al.* (2011) PLOS Comput Biol 6(1); [2] F Zanini *et al* (2017) Virus Evol 3(1).

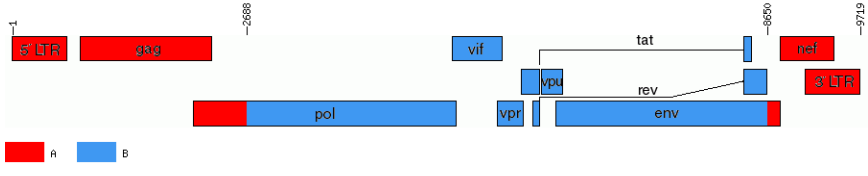
Circulating recombinant forms

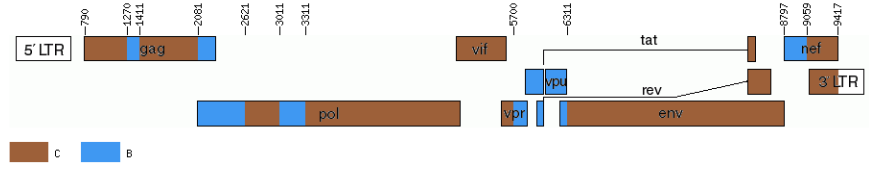


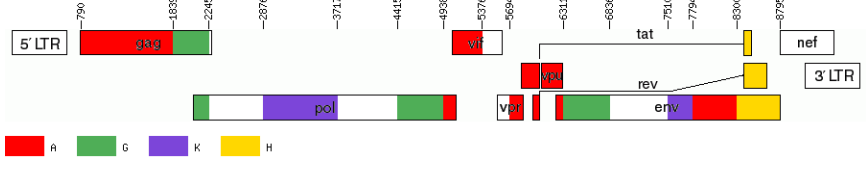


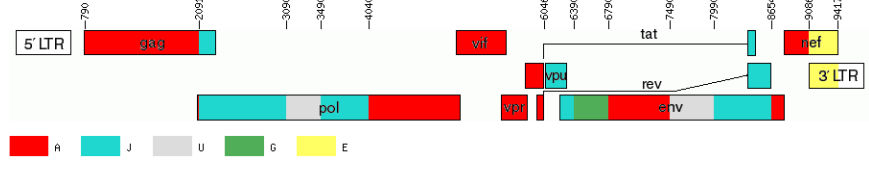






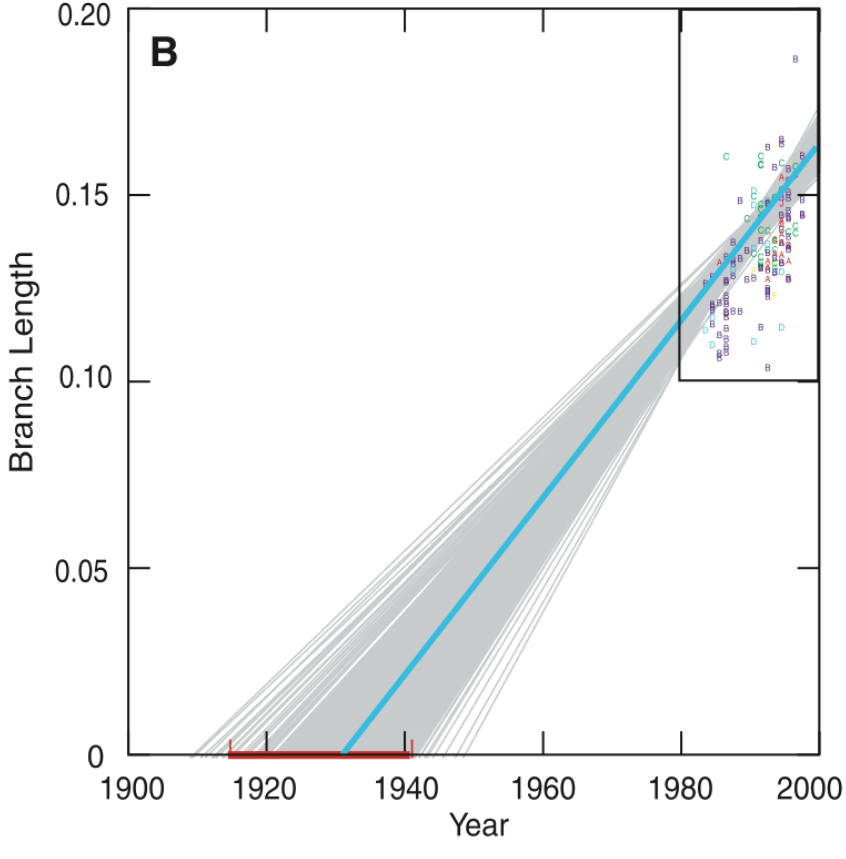






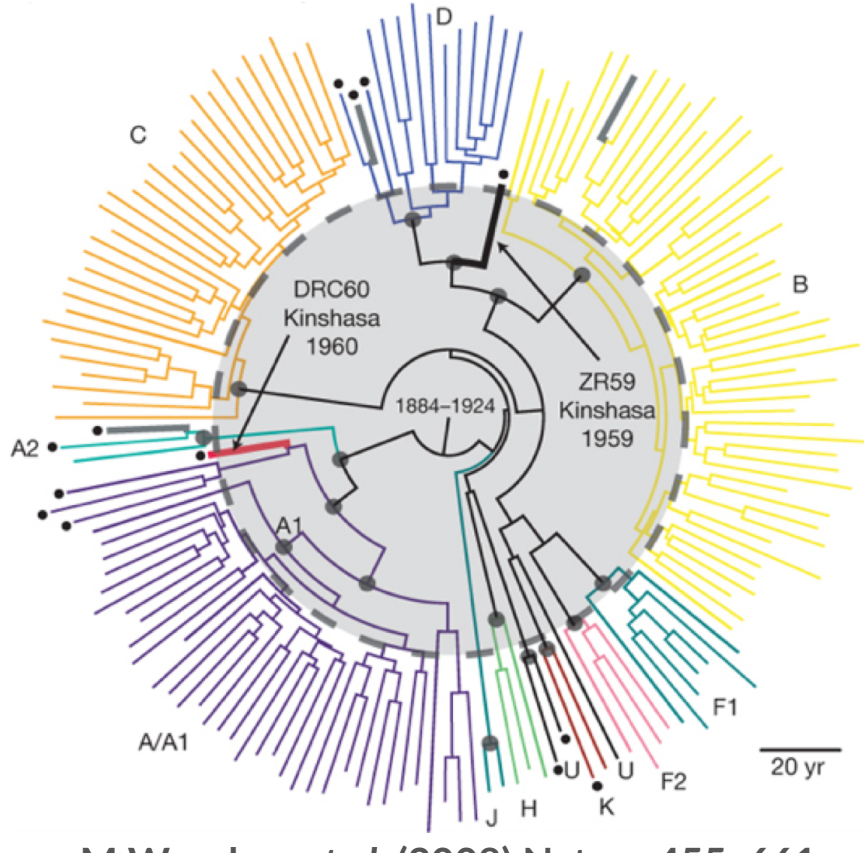
Dating the origin of HIV-1/M

Root-to-tip regression



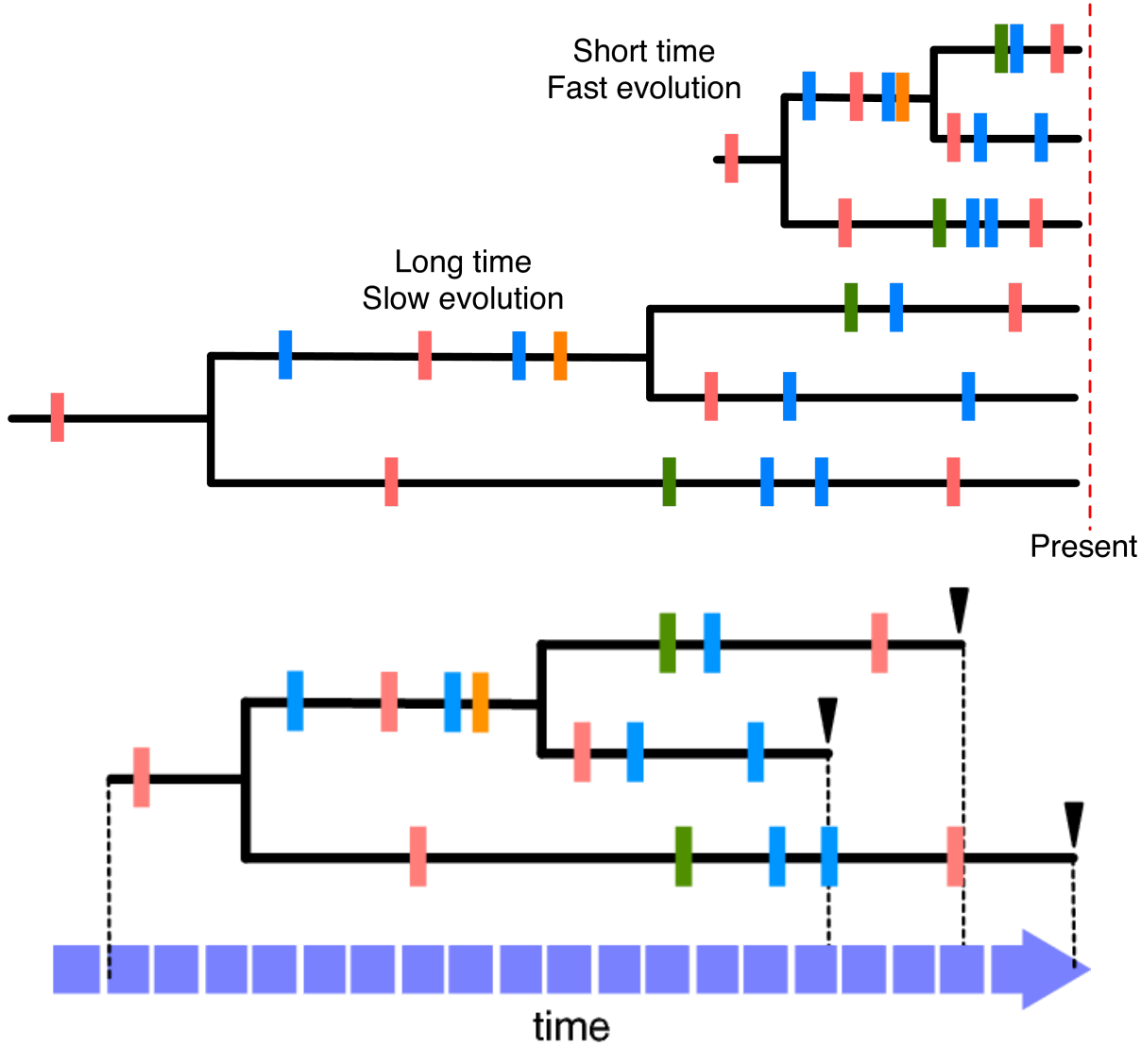
from B Korber *et al.* (2000) Science 288: 1789.

Bayesian sampling



from M Worobey *et al.* (2008) Nature 455: 661.

Molecular clock dating



More genomes

- The number of near full-length HIV-1 genomes in public databases is nearing 10,000
- A recent global consortium (PANGEA-HIV) is generating an *additional* 10,000 HIV-1 genomes sampled throughout Africa.
- An emerging consensus is that recombination is more widespread than we thought.

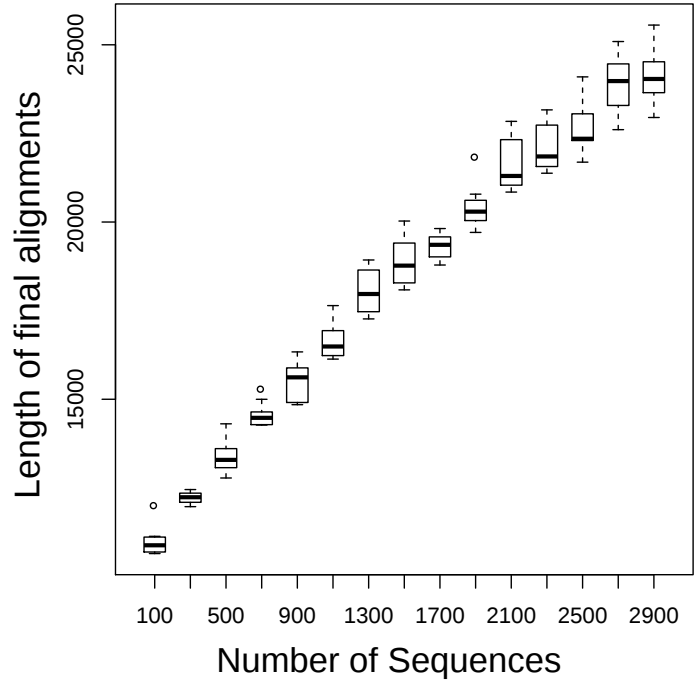
> *Given the inevitability of extensive recombination, it is highly unlikely that a single phylogeny can adequately represent the evolutionary history of HIV-1/M.*

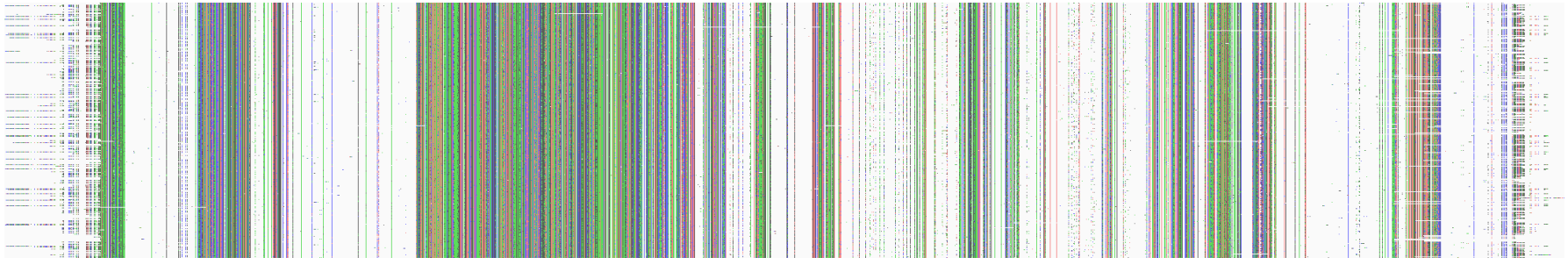
Data collection

- Queried Genbank for HIV-1 sequences of minimum length 8,000nt ($n = 7,816$)
- Manually reviewed all entries for *in vitro* clones, repeated samples, and non-group M variants.
- Reviewed associated literature for missing collection dates.
- Final total: $n = 3,900$ genome sequences.

Alignment

- An alignment is a hypothesis about the evolutionary homology of specific residues between two sequences.
- The genomic diversity of HIV-1/M causes the alignment to explode with regions of low homology.





Procrustean alignment

- Aggressive pairwise alignment of each sequence against a reference genome.
- Any insertions relative to the reference are *discarded*.
- We need the most representative reference possible!

Alignment-free clustering

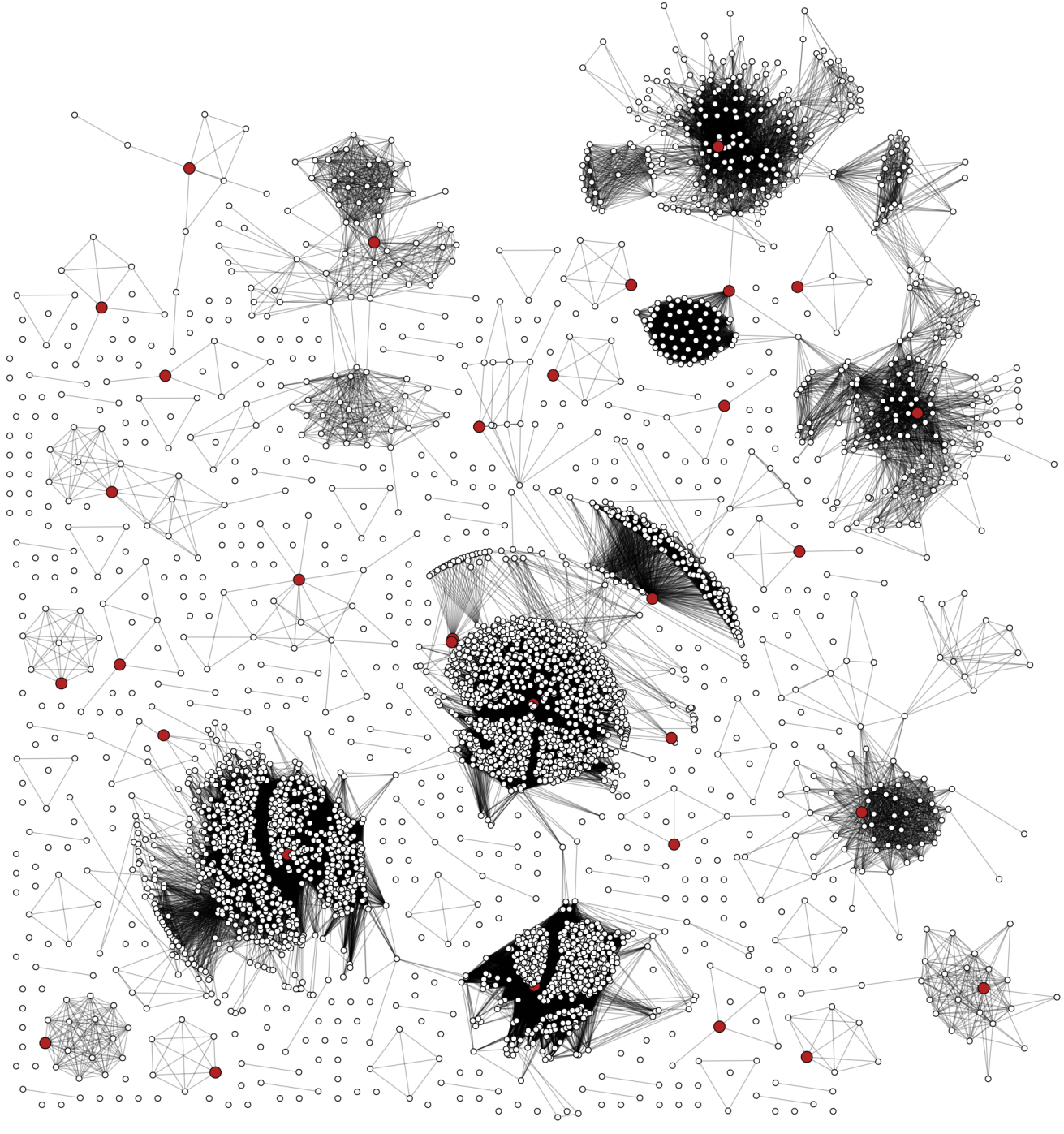
- Use a p-spectrum kernel distance to cluster genomes:

$$k(s,t) = \sum_{u \in \mathcal{A}^6} C(u,s)C(u,t)$$

where $\mathcal{A} = \{A, C, G, T\}$ and $C(x,y)$ counts occurrences of substring x in y.

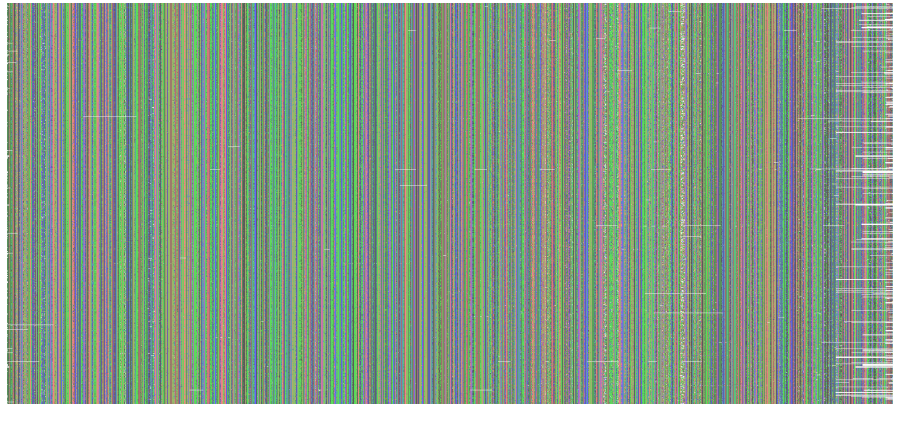
- We use the standard cosine normalization:

$$k'(s,t) = \frac{k(s,t)}{\sqrt{k(s,s)k(t,t)}}$$



Consensus sequence

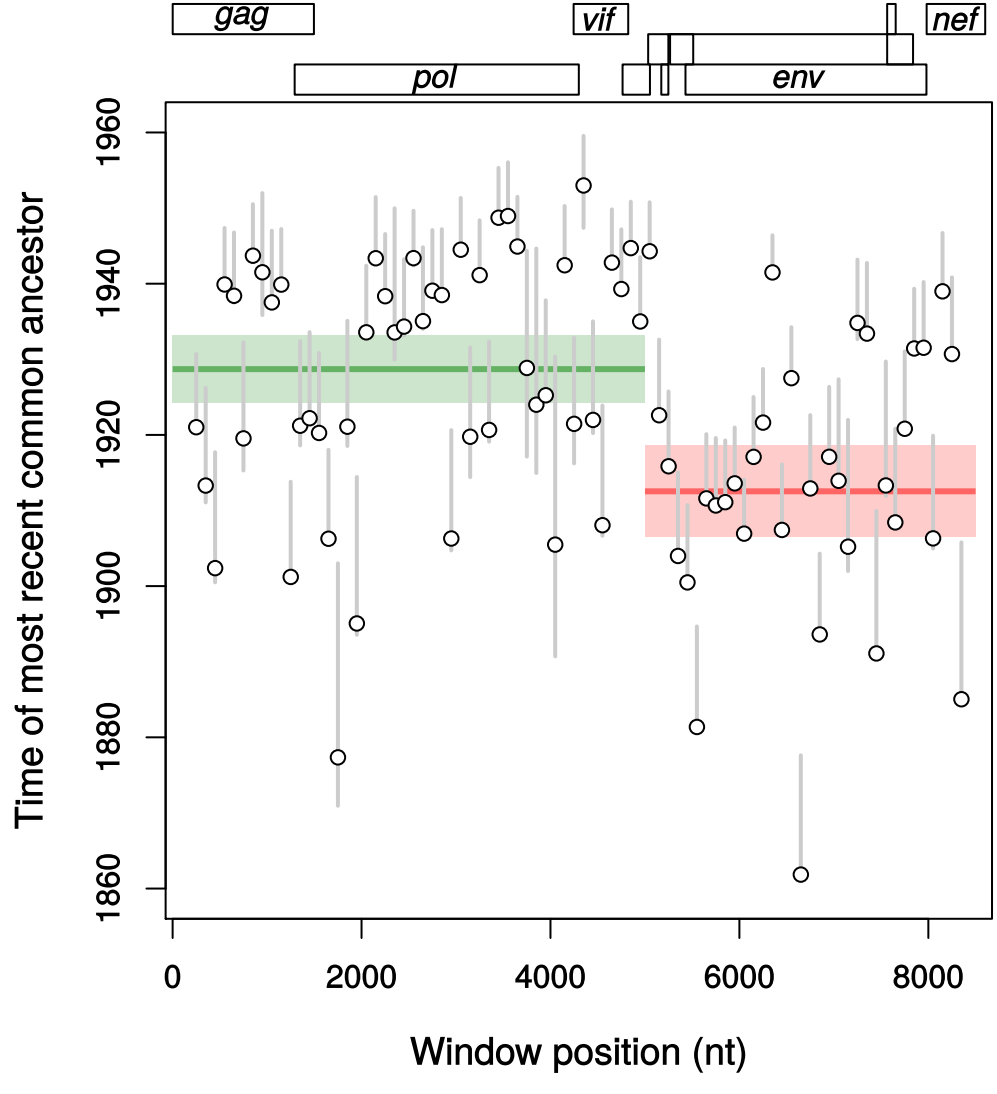
- Clustered genomes in this similarity graph with community detection algorithm (R *igraph* `cluster_leading_eigen`).
- Selected $n = 32$ centres with highest degree centrality.
- Multiple alignment of central genomes to make majority-rule consensus for Procrustean alignment.
- Final alignment 10,171nt in length.



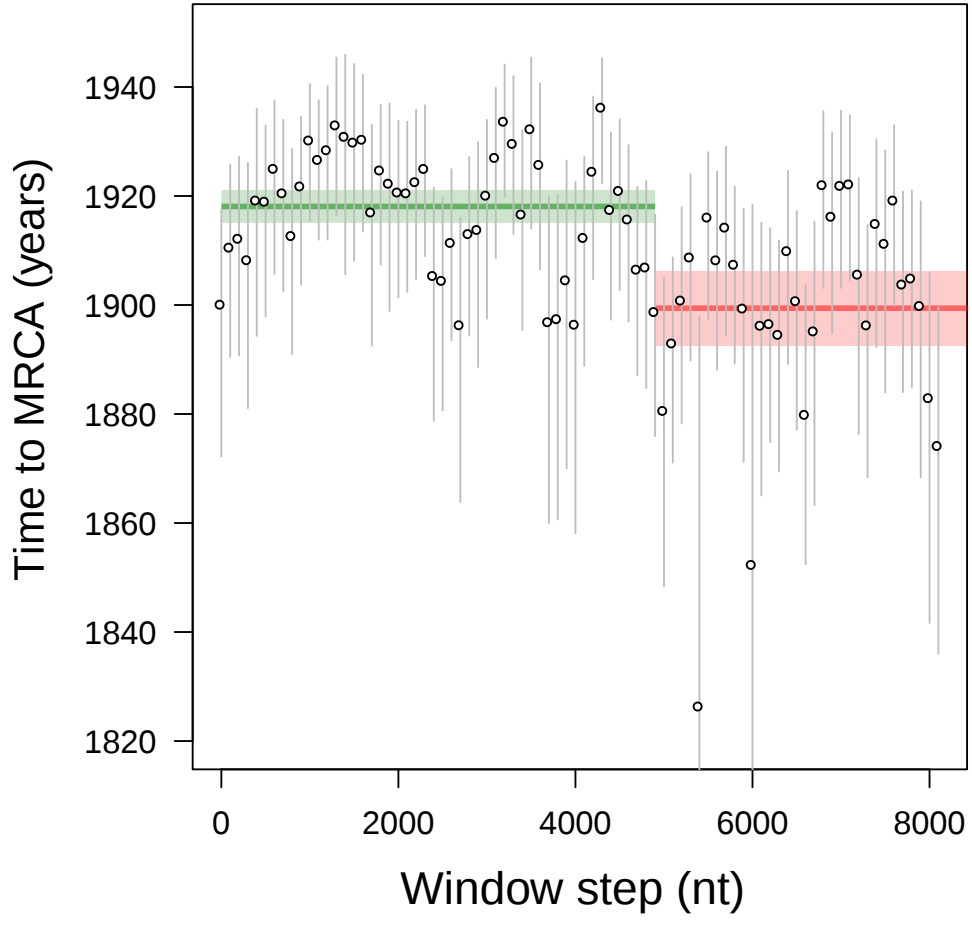
Sliding windows

- Extracted 81 windows of 500nt in steps of 100nt.
- For each window:
  - Reconstructed phylogeny by approximate ML (FastTree2).
  - Rooted phylogeny by root-to-tip regression (rtt)
  - Estimated time to the most recent common ancestor (TMRCA) under relaxed molecular clock (LSD).

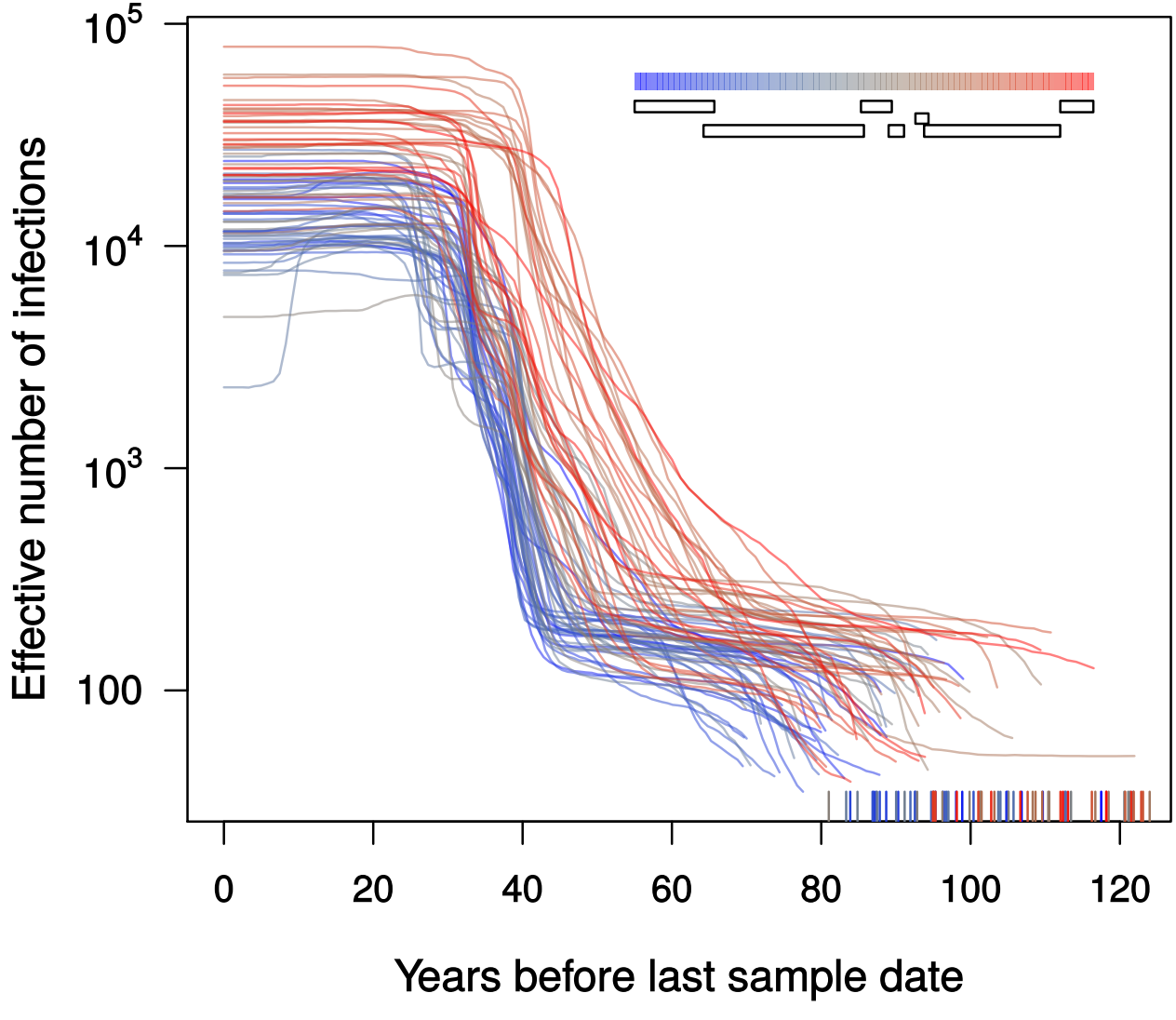
TMRCA are variable and autocorrelated along the HIV-1 genome



Similar results obtained using BEAST (data downsampled to ~300 tips).



Demographic "skyline" reconstructions from BEAST

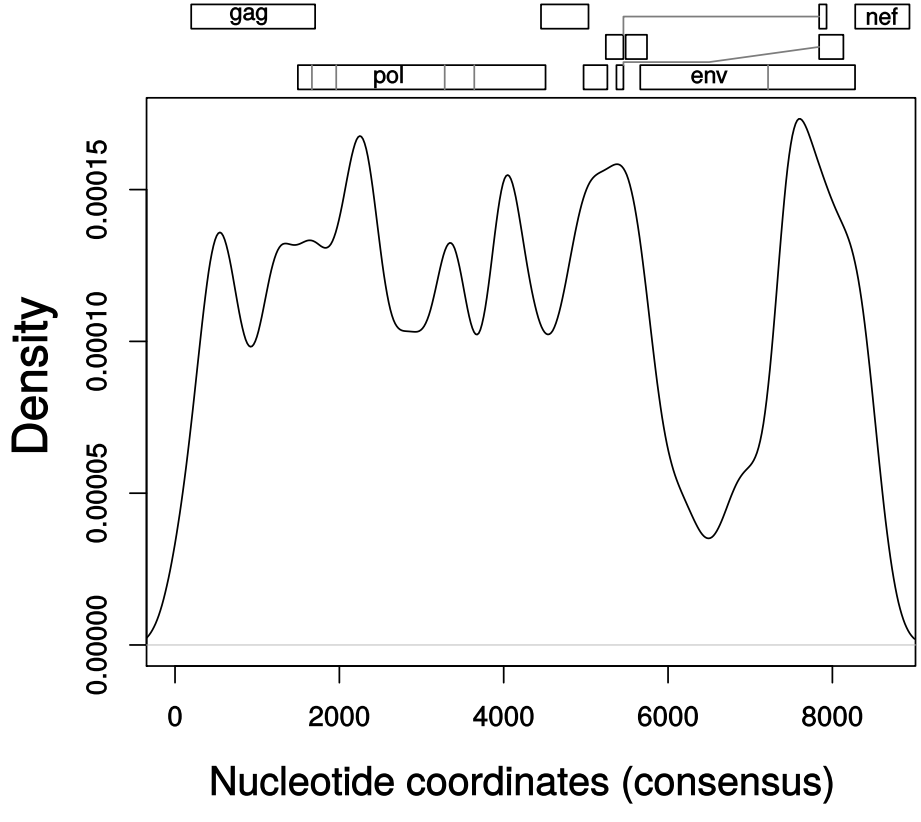


A deep recombination breakpoint?

- Why does HIV-1/M look older in *env*?
- Purifying selection may skew estimates of TMRCA.
  - No evidence of greater purifying selection in *gag* and *pol* versus *env*.
  - No evidence of saturation when comparing transitions to transversions.
- Examine signature of recombination.

Recombination analysis

- Implemented a conventional distance-based recombination detection method in Python.
- Putative breakpoints less frequent within *env*.



What does this mean?

- The deep history of HIV-1/M is recombinant.
- Is the concept of "pure" subtypes still useful?
- HIV-1 subtypes and sub-subtypes are being re-assessed.
- Selective introgression of HIV-1 *env*: first contact with the virus envelope?

A recombinant history

- How can we detect recombination if our reference points are themselves recombinant?
- Postulate 1: We cannot meaningfully describe recombinants as a mixture of "pure" subtypes on the time scale of HIV-1/M.
- Postulate 2: The recombinant history is too complex to reconstruct, *e.g.*, reticulate phylogenies.
- Postulate 3: There is enough residual homology to infer the evolutionary relationships of genome fragments.

Back to networks

- Let the homology of fragments be represented by a similarity graph.
- Graph clusters correspond to subtypes.
- As we move along the genome, fragments may shift from one cluster to another due to recombination.
- This process is analogous to a dynamic social network.

Stochastic blockmodeling

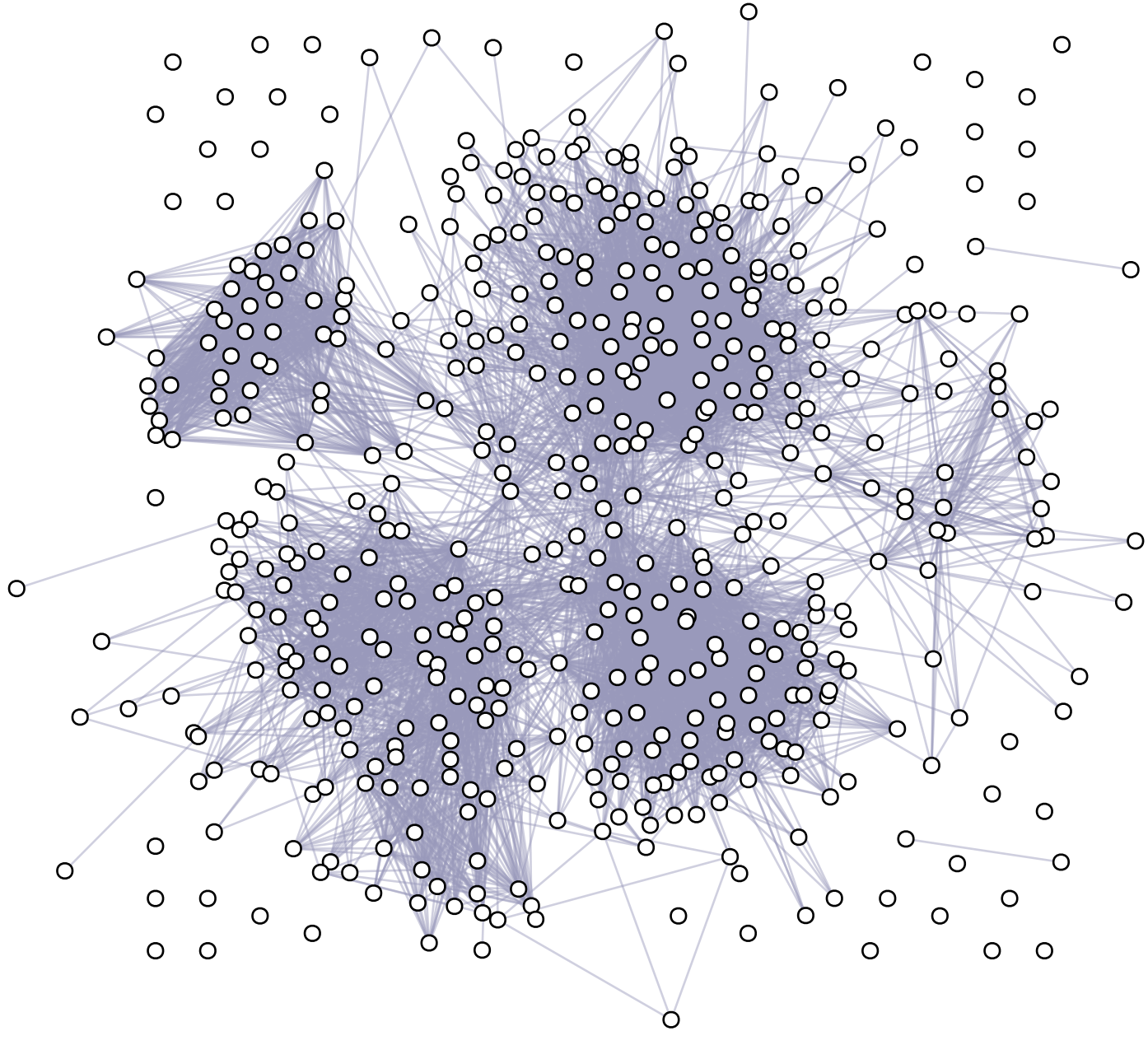
- Generative models for network communities.
- Each node is assigned to one of K communities (blocks) — block membership is a latent (unobservable) state of the node.
- Different edge probabilities within and between blocks.
- Transition of nodes among blocks over time is described by a Markov chain.

Data reduction

- SBM becomes computationally non-feasible with the complete data set.
- We reduced the $n = 3,900$ genome sequences to the most diverse $n = 500$.
- Progressive "pruning" of the shortest branches from the ML phylogeny.

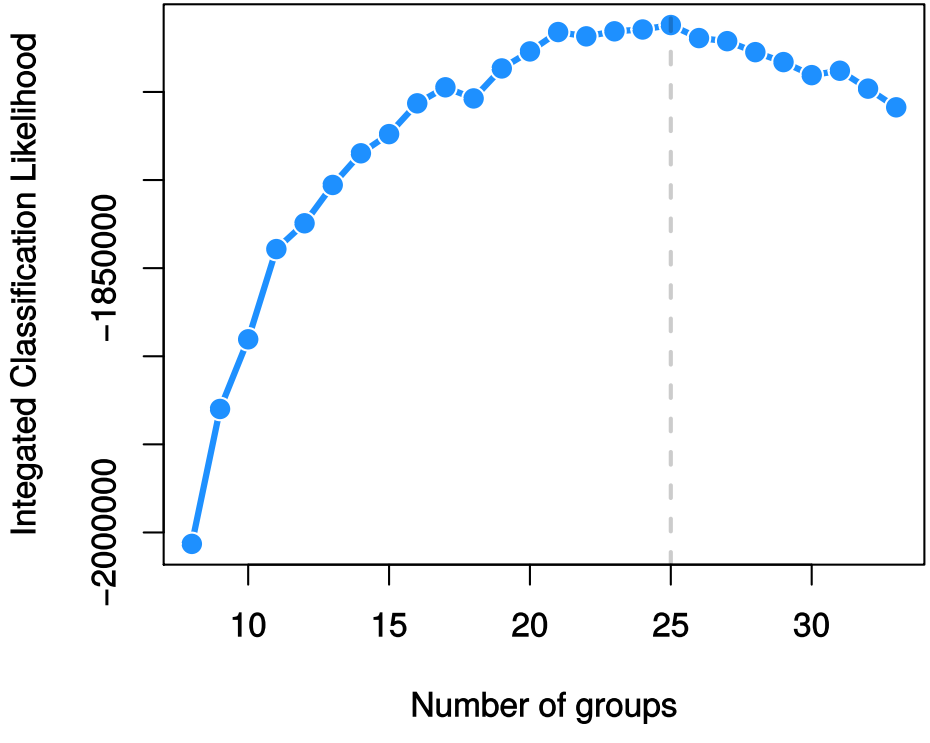
Constructing a graph

- We calculated the Tamura-Nei (1993; TN93) genetic distance for every pair of fragments in a window.
- Used the lower 10% quantile of the TN93 distribution as the graph-defining threshold for each window.
- Processed the series of graphs along the HIV-1 genome with dynamic stochastic block models (R package dynsbm; Matias and Miele).

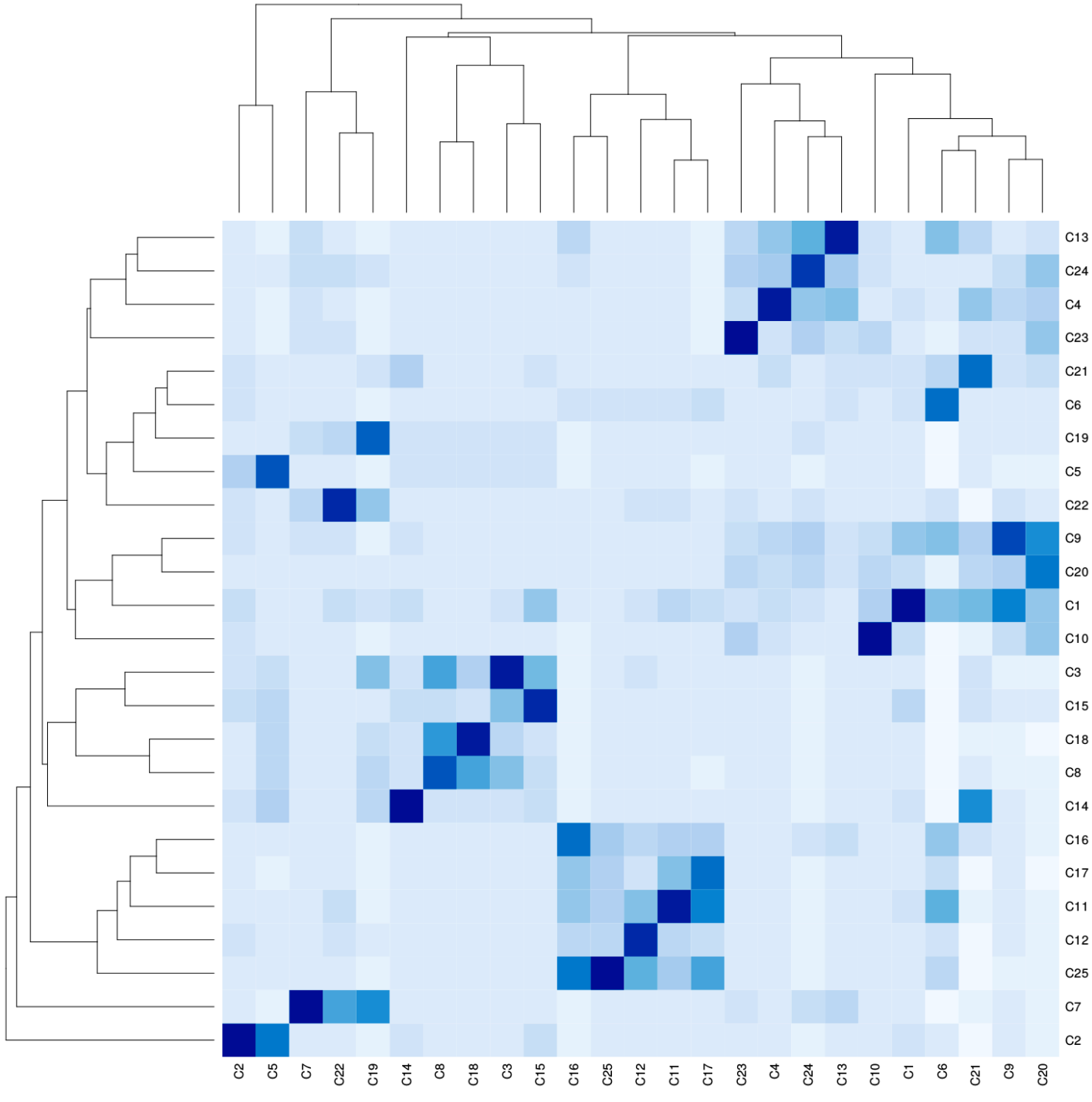


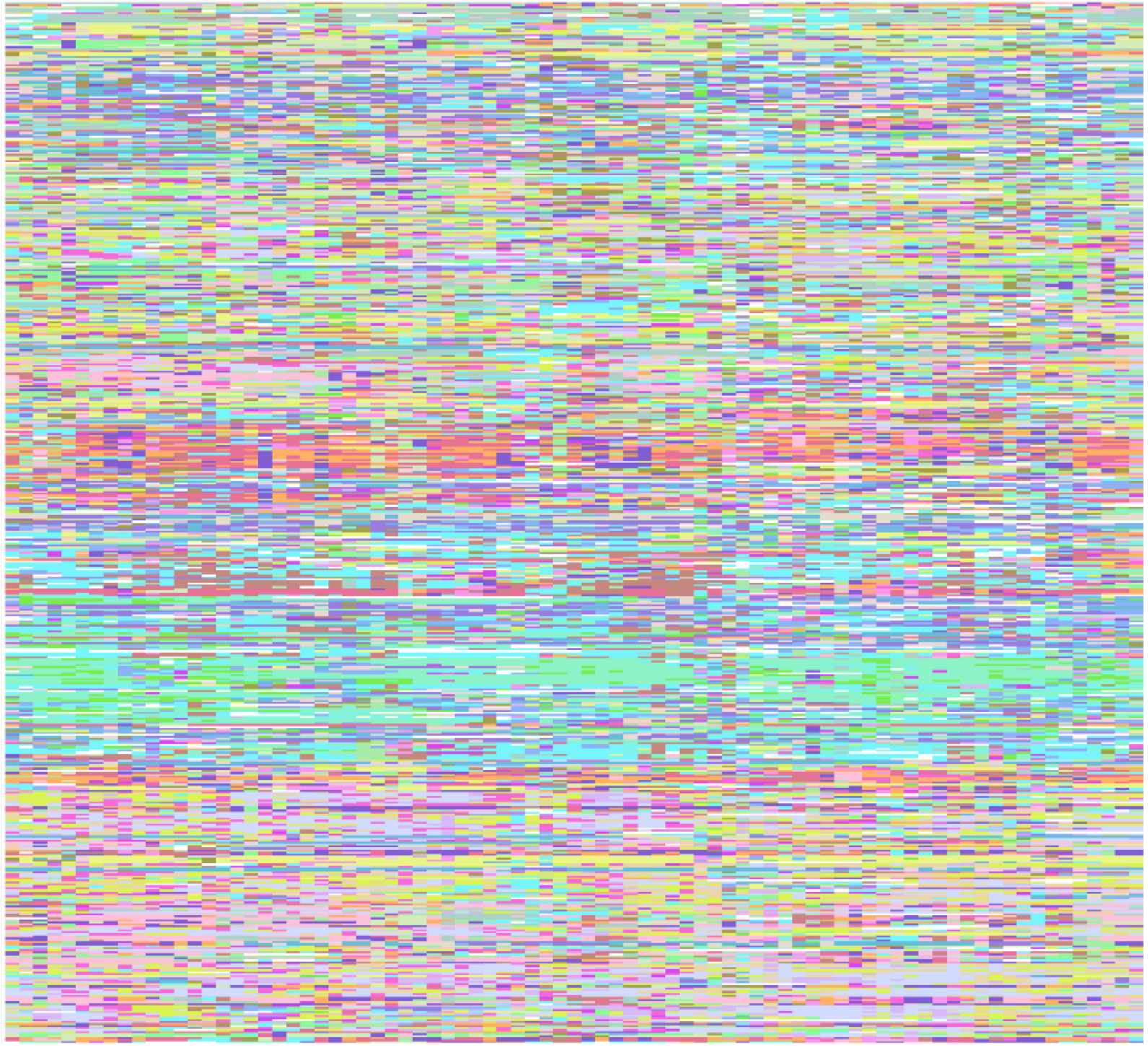
Optimal K=25

- Integrated classification likelihood (ICL) criterion avoids overestimation of the number of clusters by penalizing likelihood[1].

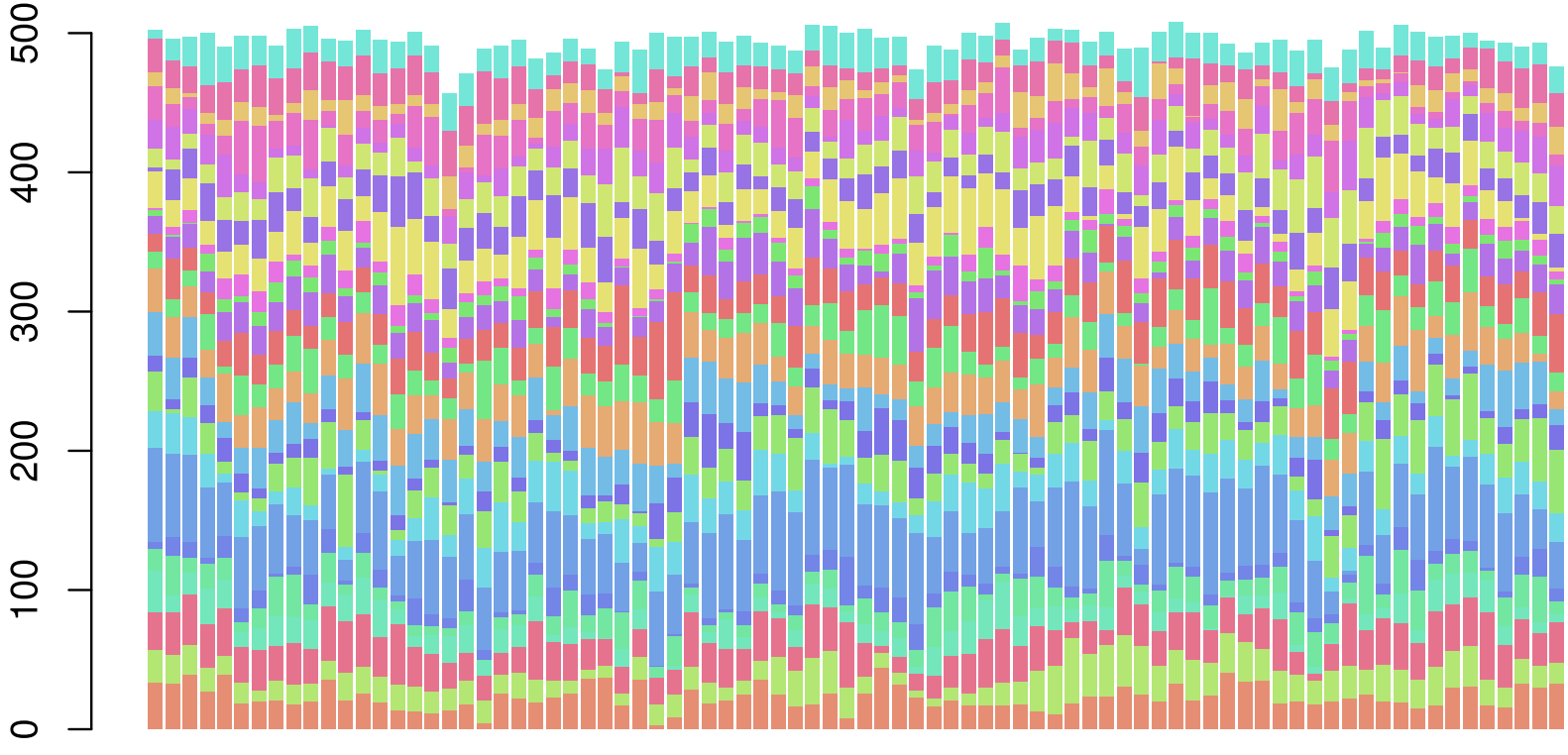


[1] C Biernacki *et al.* (1998) Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. INRIA.

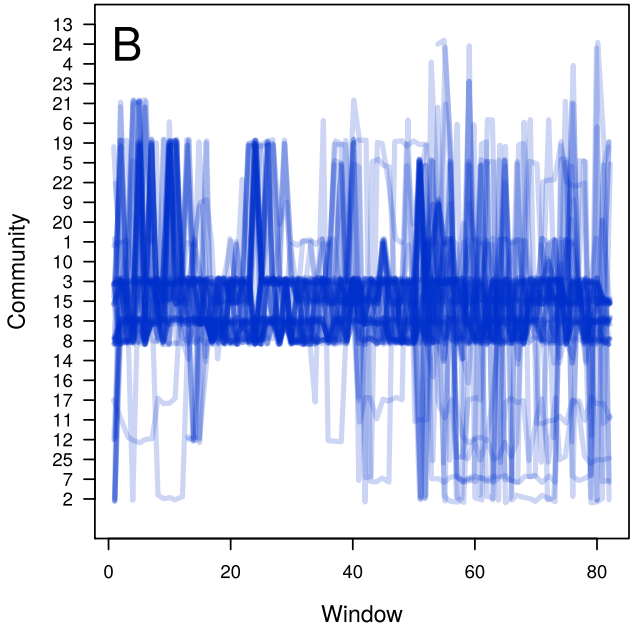

C13
C24
C4
C23
C21
C6
C19
C5
C22
C9
C20
C1
C10
C3
C15
C18
C8
C14
C16
C17
C11
C12
C25
C7
C2
C2
C5
C7
C22
C19
C14
C8
C18
C3
C15
C16
C25
C12
C11
C17
C23
C4
C24
C13
C10
C1
C6
C21
C9
C20

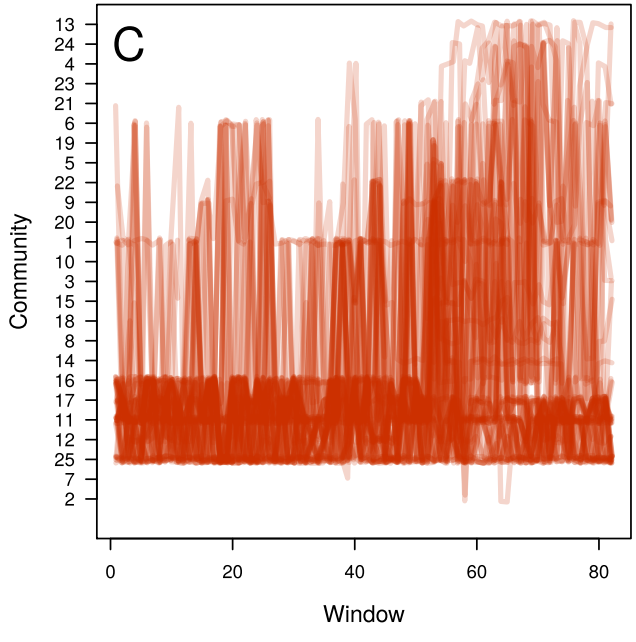


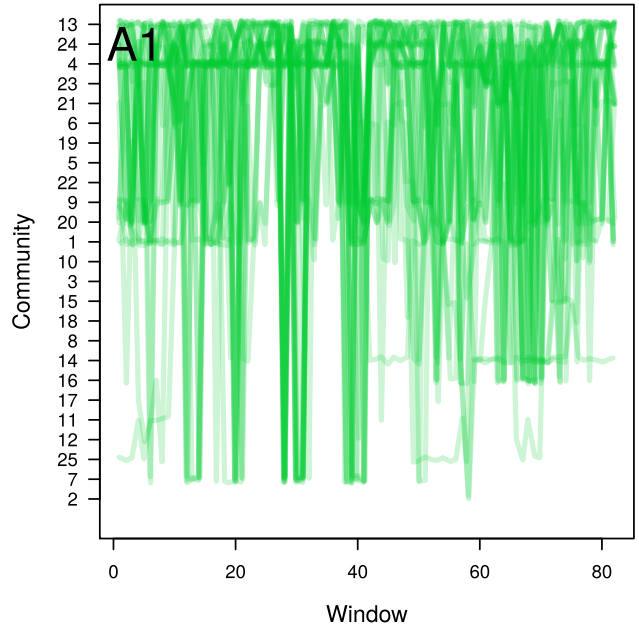
Communities are fairly stable



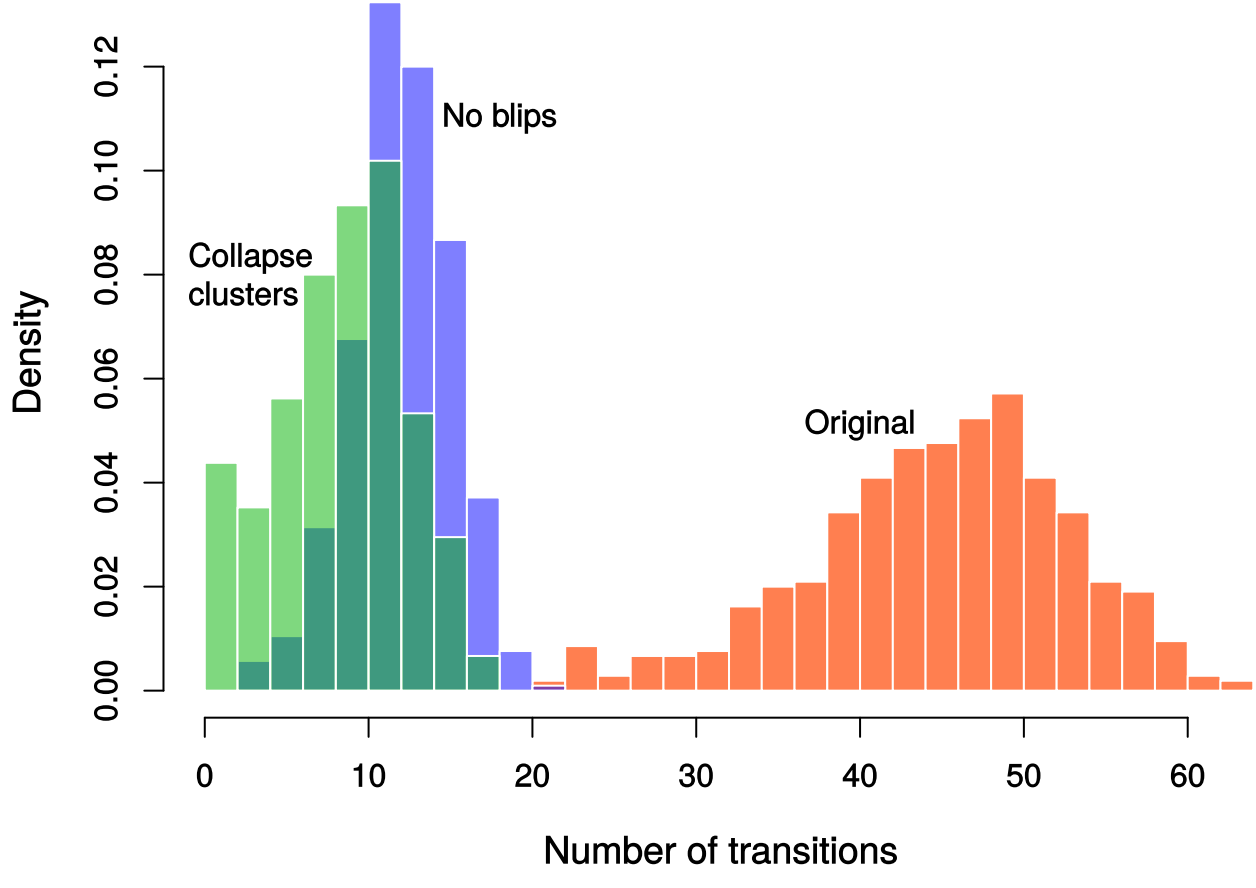
Communities correlate with major subtypes



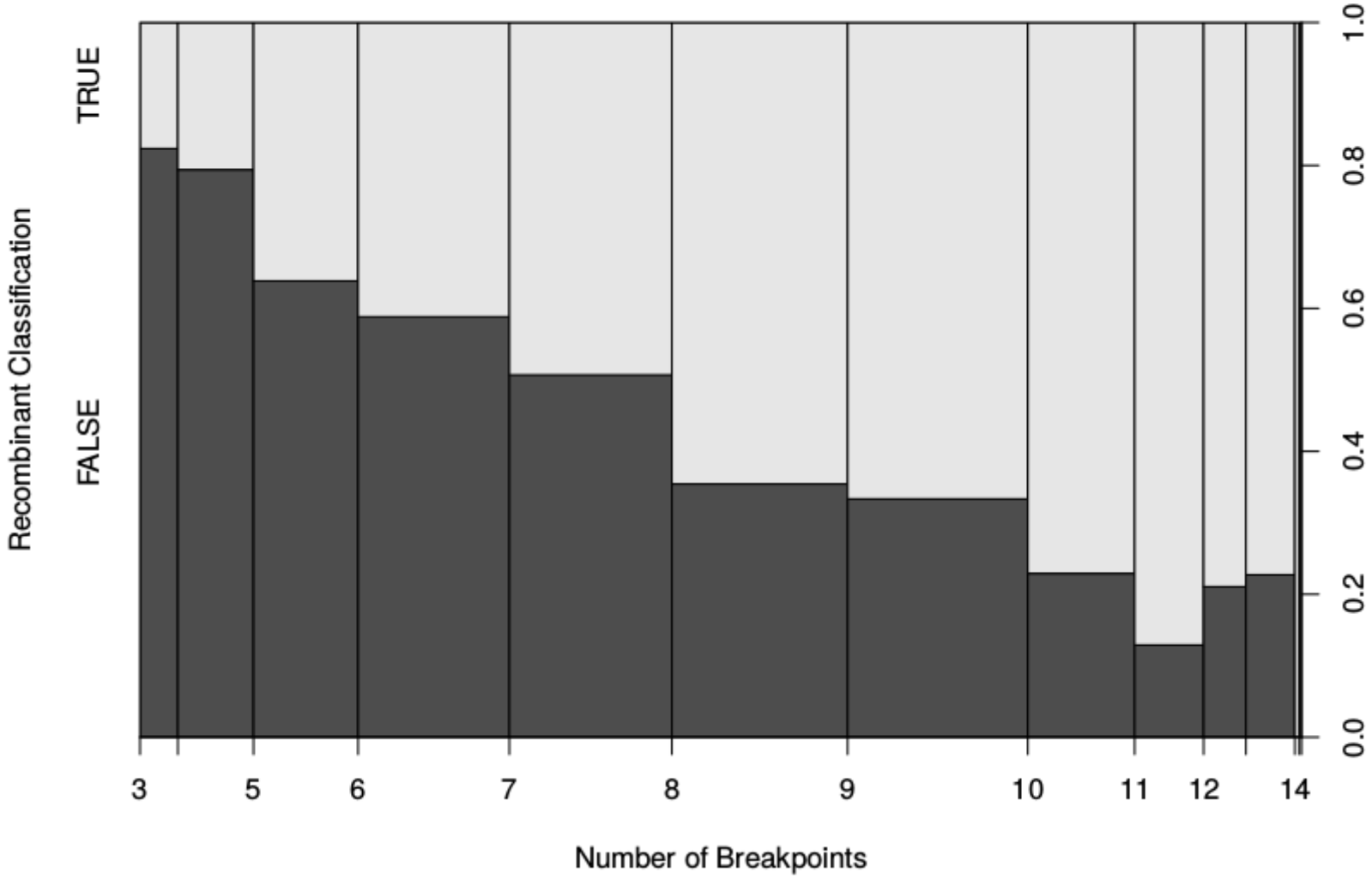




Number of putative breakpoints



Community transitions are consistent with "known" recombinants.



Plot excludes "blips" (transitions from A to B immediately followed by a revision from B to A in the next window).

A new nomenclature proposal?

- Unsupervised clustering of genomic variation may define new centres of diversity.
- Can we select new reference points for characterizing recent recombinants?
- Need simulation experiments to assess what network communities really mean.

Left to right: Dr. Mariano Avino, **Dr. Abayomi Olabode**, Garway Ng, Lisa-Monique Edward, John Palmer, Connor Chato, *Art and Fiona*, David Dick, Laura Muñoz-Baena. (Photo credit: Marin Poon.)

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada