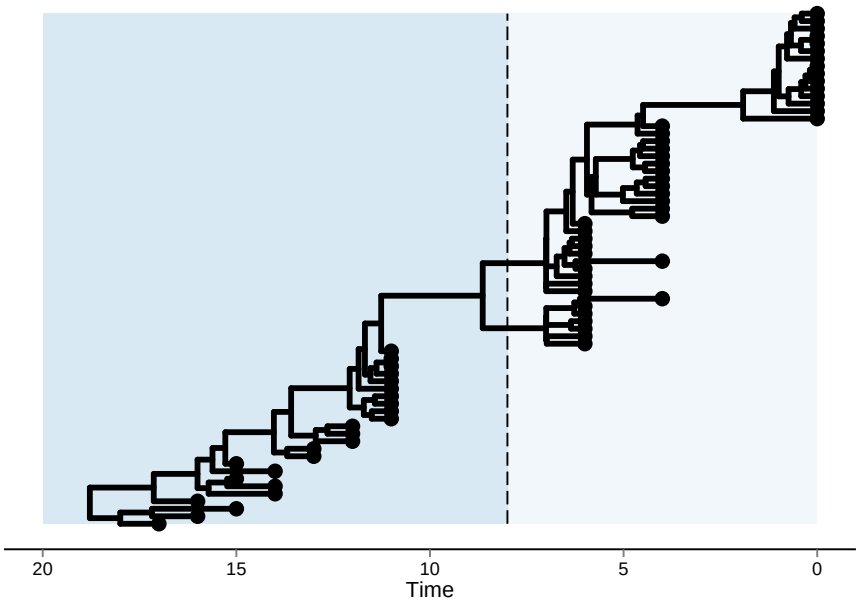
Inferring large-scale heterogeneous evolutionary processes through time

Filip Bielejec

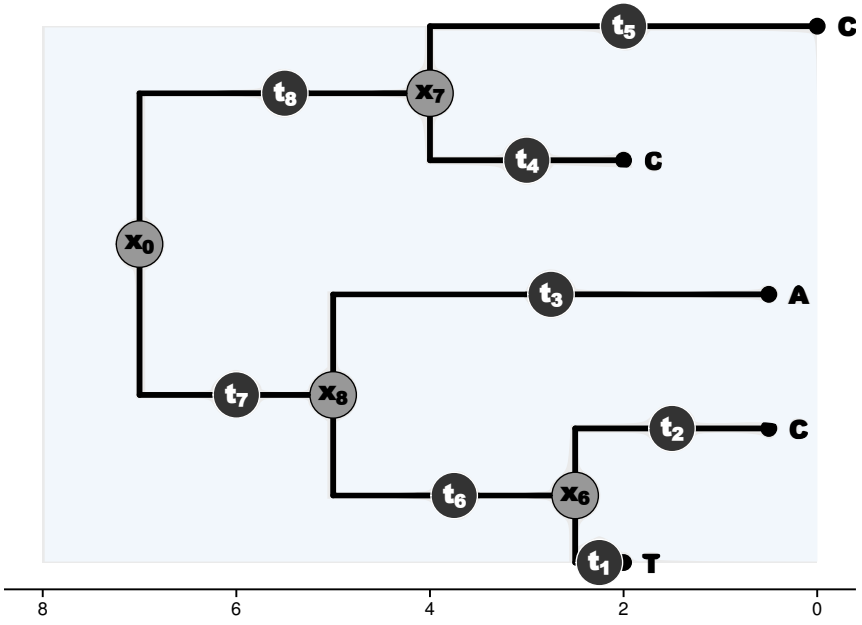
Evolutionary and Computational Virology
Rega Institute
Department of Microbiology and Immunology
KU Leuven, Belgium.
www.kuleuven.be/rega/ecv

Epoch substitution model concept



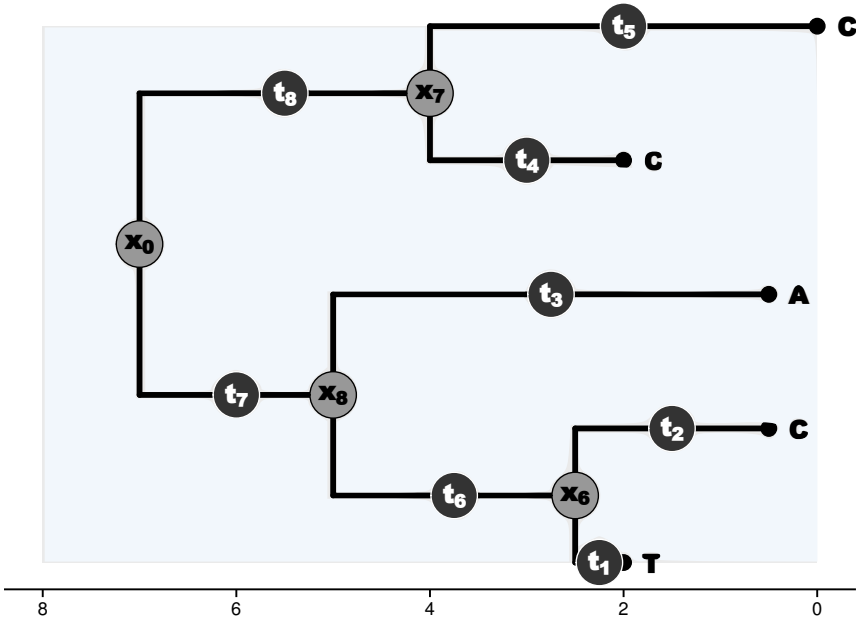
- Imagine some event in time causes a change in substitution process.
- We have a rough estimate of when the transition time has occured.
- We collect the observed data, put forward an evolutionary model for each epoch and calculate the likelihood.

Calculating likelihood on tree - homogenous case



- Assumptions of independence:
$$\sum_{X_0 \in S} \sum_{X_6 \in S} \sum_{X_7 \in S} \sum_{X_8 \in S} [\pi_{x_0} \cdot p_{x_0 x_8}(t_7) \cdot p_{x_0 x_7}(t_8) \cdot p_{x_8 x_6}(t_6) \cdot p_{x_8 A}(t_3) \cdot p_{x_7 C}(t_4) \cdot p_{x_7 C}(t_5) \cdot p_{x_6 T}(t_1) \cdot p_{x_6 C}(t_2)]$$
- Inside brackets: probability of the data $TCACC$ for the tips and x_0, x_6, x_7, x_8 for the ancestral nodes.
- Outside brackets: Integrating out the unobserved data over discrete state space S.

Calculating likelihood on tree - homogenous case



- Assumptions of independence:
$$\sum_{X_0 \in S} \sum_{X_6 \in S} \sum_{X_7 \in S} \sum_{X_8 \in S} [\pi_{x_0} \cdot p_{x_0 x_8}(t_7) \cdot p_{x_0 x_7}(t_8) \cdot p_{x_8 x_6}(t_6) \cdot p_{x_8 A}(t_3) \cdot p_{x_7 C}(t_4) \cdot p_{x_7 C}(t_5) \cdot p_{x_6 T}(t_1) \cdot p_{x_6 C}(t_2)]$$
- **Inside brackets:** probability of the data $TCACC$ for the tips and x_0, x_6, x_7, x_8 for the ancestral nodes.
- **Outside brackets:** Integrating out the unobserved data over discrete state space S.

Felsensteins tree pruning (Felsenstein 1981)

$$L_i(x_i) = \left[\sum_{x_j \in S} p_{x_i x_j}(t_j) L_j(x_j)\right] \times \left[\sum_{x_k \in S} p_{x_i x_k}(t_k) L_k(x_k)\right]$$

- Evolutionary model quantifies transition probabilities $p_{x_i x_j}(t_j)$.
- Probability $L_i(x_i)$ of observing data at the descendant tips of node i given state x_i at node i conveniently expressed in terms of probabilities at nodes j and k.
- Requires post-order tree traversal.

Felsensteins tree pruning (Felsenstein 1981)

$$L_i(x_i) = \left[\sum_{x_j \in S} p_{x_i x_j}(t_j) L_j(x_j)\right] \times \left[\sum_{x_k \in S} p_{x_i x_k}(t_k) L_k(x_k)\right]$$

- Evolutionary model quantifies transition probabilities $p_{x_i x_j}(t_j)$.
- Probability $L_i(x_i)$ of observing data at the descendant tips of node i given state x_i at node i conveniently expressed in terms of probabilities at nodes j and k.
- Requires post-order tree traversal.

Felsensteins tree pruning (Felsenstein 1981)

$$L_i(x_i) = \left[\sum_{x_j \in S} p_{x_i x_j}(t_j) L_j(x_j)\right] \times \left[\sum_{x_k \in S} p_{x_i x_k}(t_k) L_k(x_k)\right]$$

- Evolutionary model quantifies transition probabilities $p_{x_i x_j}(t_j)$.
- Probability $L_i(x_i)$ of observing data at the descendant tips of node i given state x_i at node i conveniently expressed in terms of probabilities at nodes j and k.
- Requires post-order tree traversal.

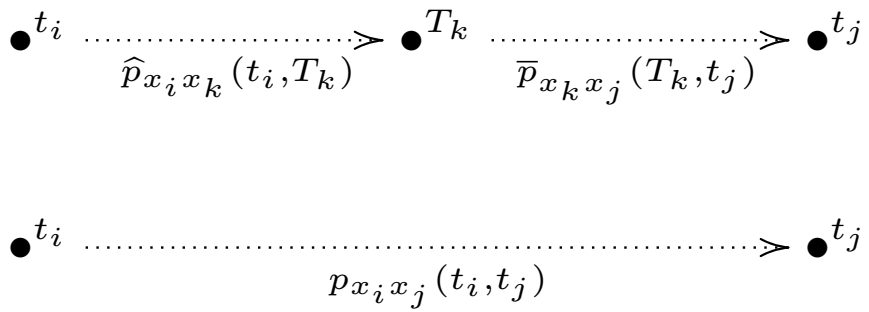
Time homogeneity

- Central to the recursive pruning algorithm is the specification of branch specific transition probabilities over time t_i to t_j: $p_{x_i x_j}(t_j, t_j + t_i)$
- Time homogeneity assumption allows us to express them in terms of just the branch length: $p_{x_i x_j}(t_i, t_j + t_i) = p_{x_i x_j}(0, t_j) = p_{x_i x_j}(t_j)$
- In the epoch model specification, if particular branch spans over a transition time $T_k,\ k \in \{0, \cdots, M\}$ the homogeneity assumption is violated.

Time homogeneity

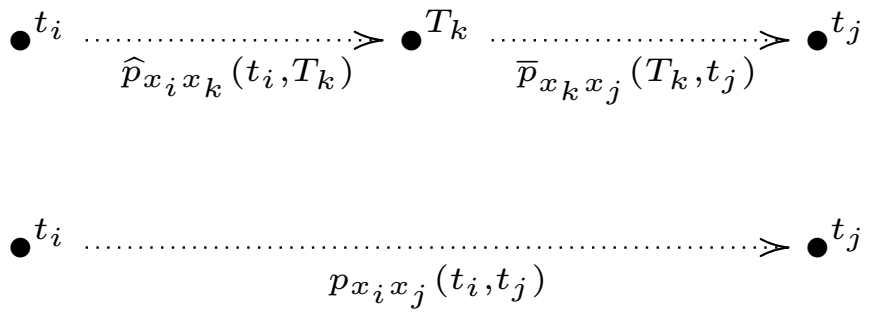
- Central to the recursive pruning algorithm is the specification of branch specific transition probabilities over time t_i to t_j: $p_{x_i x_j}(t_j, t_j + t_i)$
- Time homogeneity assumption allows us to express them in terms of just the branch length: $p_{x_i x_j}(t_i, t_j + t_i) = p_{x_i x_j}(0, t_j) = p_{x_i x_j}(t_j)$
- In the epoch model specification, if particular branch spans over a transition time $T_k,\ k \in \{0, \cdots, M\}$ the homogeneity assumption is violated.

Relaxing time homogeneity



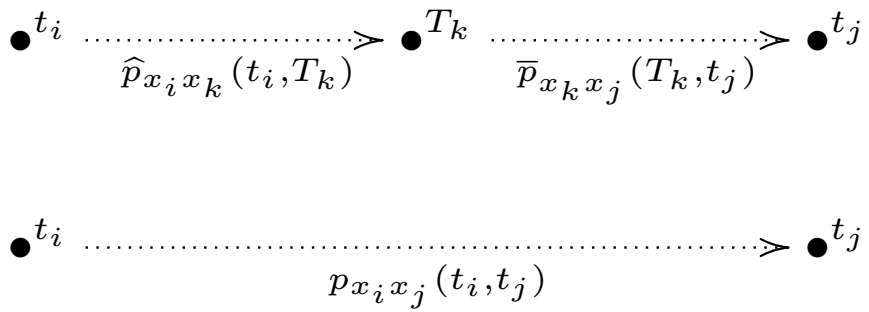
- To handle the discontinuity we numerically integrate out the unobserved state x_k at the transition time T_k:
- $p_{x_i x_j}(t_i, t_j) = \sum_{x_k \in S} \overline{p}_{x_i x_k}(t_j - T_k) \cdot \widehat{p}_{x_k x_j}(T_k - t_j)$
- Equivalent to multiplying transition matrices: $\mathbf{P}(t_j - t_i) = \overline{\mathbf{P}}(t_j - T_k) \times \widehat{\mathbf{P}}(T_k - t_i)$

Relaxing time homogeneity



- To handle the discontinuity we numerically integrate out the unobserved state x_k at the transition time T_k:
- $p_{x_i x_j}(t_i, t_j) = \sum_{x_k \in S} \overline{p}_{x_i x_k}(t_j - T_k) \cdot \widehat{p}_{x_k x_j}(T_k - t_j)$
- Equivalent to multiplying transition matrices: $\mathbf{P}(t_j - t_i) = \overline{\mathbf{P}}(t_j - T_k) \times \widehat{\mathbf{P}}(T_k - t_i)$

Relaxing time homogeneity



- To handle the discontinuity we numerically integrate out the unobserved state x_k at the transition time T_k:
- $p_{x_i x_j}(t_i, t_j) = \sum_{x_k \in S} \overline{p}_{x_i x_k}(t_j - T_k) \cdot \widehat{p}_{x_k x_j}(T_k - t_j)$
- Equivalent to multiplying transition matrices: $\mathbf{P}(t_j - t_i) = \overline{\mathbf{P}}(t_j - T_k) \times \widehat{\mathbf{P}}(T_k - t_i)$

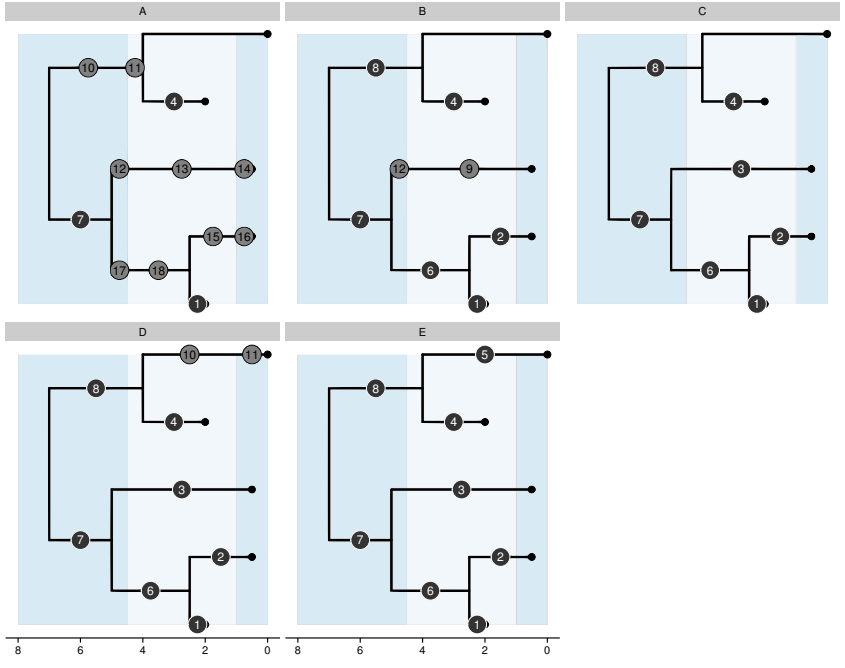
Computational burden

- Rooted tree topology $\mathbf{F}$ with N tips, a character sequence of length L, obtaining K distinct values, site rate variation with C categories.
- Likelihood computation *via* pruning is $\mathcal{O}(K^2 \times N \times C \times L)$.
- Each matrix multiplicaton is $\mathcal{O}(K^3)$. Upper boundary on the number of operations is $(2N - 2) \times (S - 1)$, where S is the number of substitution processes operating on $\mathbf{F}$.

Fine grain parallelism

- High performance kernel for matrix multiplication on GPU devices implemented as part of BEAGLE library (Suchard and Rambaut 2009, Ayres *et al.* 2012).
- Divide & conquer strategy - each memory entry in product matrix is computed by single thread:
- $\{\mathbf{P}\}_{ij} = \sum_{k \leq \frac{K}{BLOCK\_SIZE}} \left\{\overline{\mathbf{P}}\right\}_{ik} \times \left\{\widehat{\mathbf{P}}\right\}_{kj}$

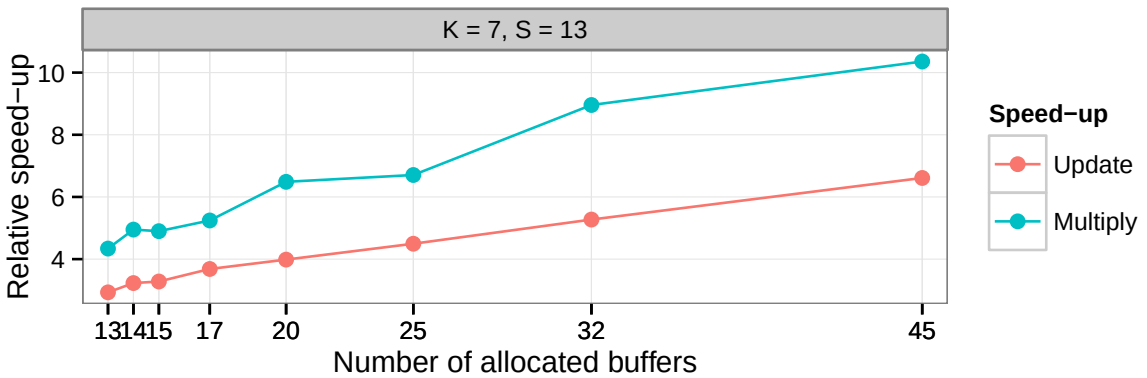
Coarse grain parallelism

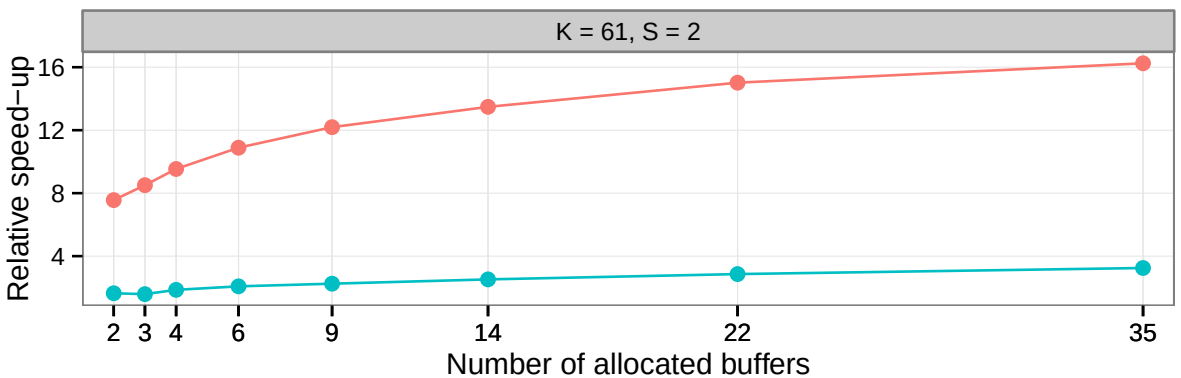


- GPUs utilize host-device programming model.
- Front end update-multiply routine implemented in BEAST (Drummond *et al.* 2012) collects independent operations into separate queues and sends them for asynchronous execution.
- Performance is highly dependent on the queue size.

Coarse grain parallelism

- Memory-speed tradeoff.
- Figure presents the GPU speed-ups induced by different choices of queue size.
- K denotes state space size, S denotes number of substitution processes.





Within-host HIV evolutionary dynamics

- We re-analyze within-host HIV-1 sequence data from eight patients sampled for 6 to 12 years until developing AIDS (Shankarappa *et al.* 1999).
- Prior analysis show consistent pattern of divergence stabilization at late stage infection.
- **Immune relaxation**: the damage in hosts immune system leads to reduced selective pressure on the virus (impacting non-synonymous rates of substitutions only).
- **Cellular exhaustion**: decreased availability of target cells in late stage infection provides less opportunity for viral replication (reducing both synonymous and non-synonymous rates of substitutions).

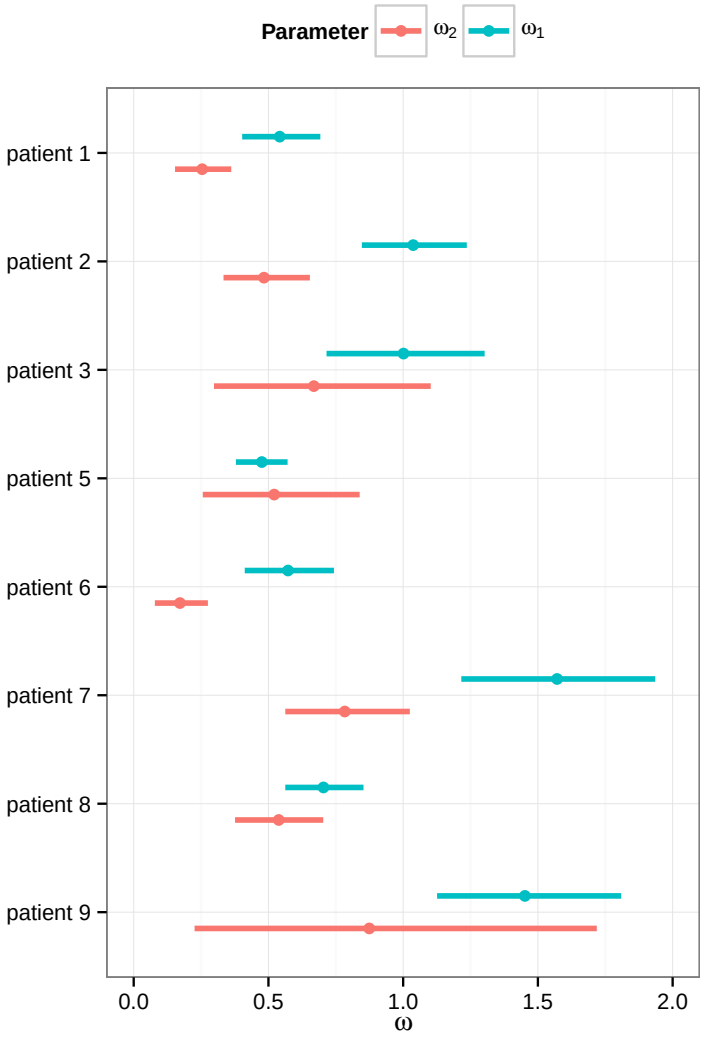
Within-host HIV evolutionary dynamics

- We re-analyze within-host HIV-1 sequence data from eight patients sampled for 6 to 12 years until developing AIDS (Shankarappa *et al.* 1999).
- Prior analysis show consistent pattern of divergence stabilization at late stage infection.
- **Immune relaxation**: the damage in hosts immune system leads to reduced selective pressure on the virus (impacting non-synonymous rates of substitutions only).
- **Cellular exhaustion**: decreased availability of target cells in late stage infection provides less opportunity for viral replication (reducing both synonymous and non-synonymous rates of substitutions).

Within-host HIV evolutionary dynamics

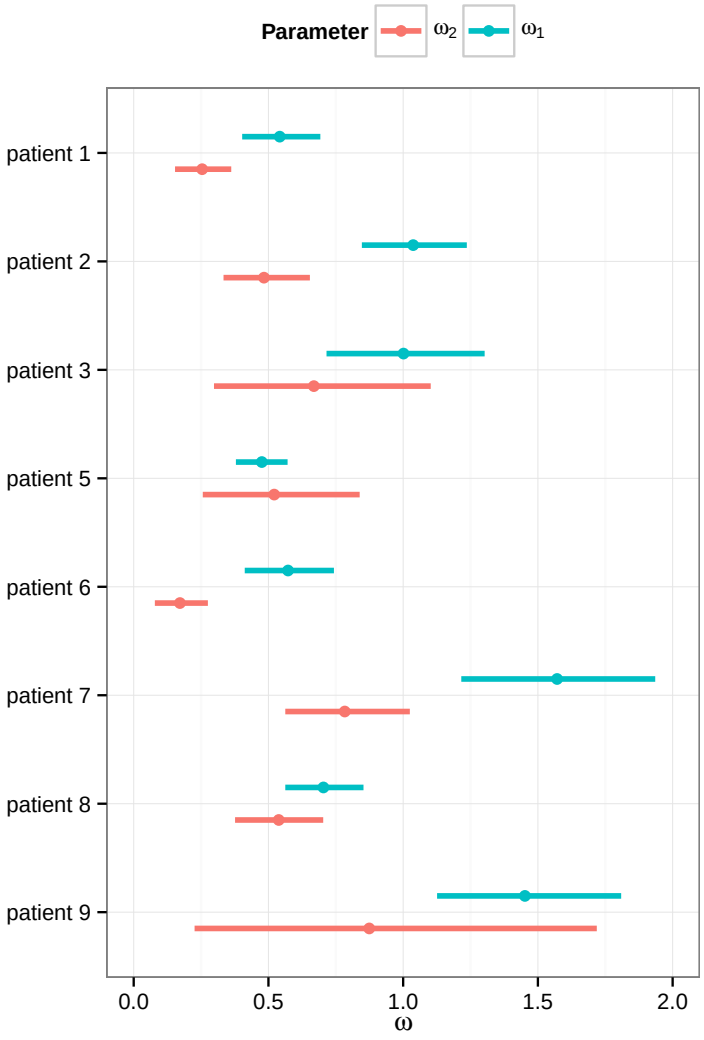
- We re-analyze within-host HIV-1 sequence data from eight patients sampled for 6 to 12 years until developing AIDS (Shankarappa *et al.* 1999).
- Prior analysis show consistent pattern of divergence stabilization at late stage infection.
- **Immune relaxation**: the damage in hosts immune system leads to reduced selective pressure on the virus (impacting non-synonymous rates of substitutions only).
- **Cellular exhaustion**: decreased availability of target cells in late stage infection provides less opportunity for viral replication (reducing both synonymous and non-synonymous rates of substitutions).

Within-host HIV evolutionary dynamics



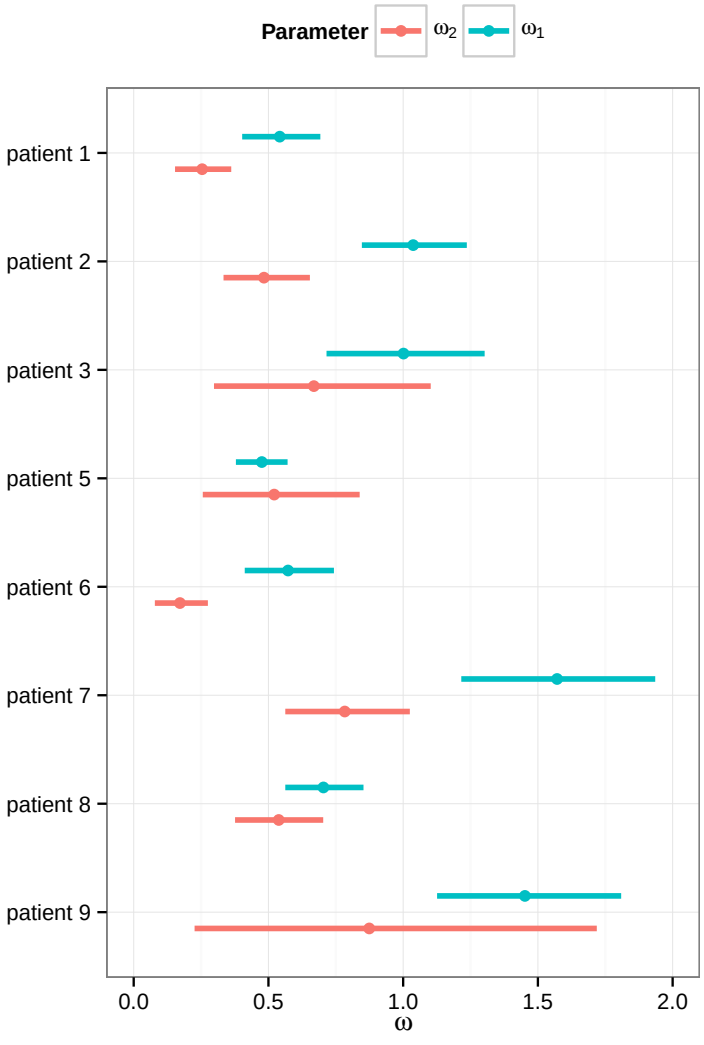
- Two-epoch discretization with separate GY94 (Goldman and Yang 1994) codon substitution models.
- Parameters of interest: ω_1 (dN/dS before progression time T_1) and ω_2 (dN/dS after progression).
- HPD intervals show a general decrease in parameter estimates after progression time.

Within-host HIV evolutionary dynamics



- Two-epoch discretization with separate GY94 (Goldman and Yang 1994) codon substitution models.
- Parameters of interest: ω_1 (dN/dS before progression time T_1) and ω_2 (dN/dS after progression).
- HPD intervals show a general decrease in parameter estimates after progression time.

Within-host HIV evolutionary dynamics



- Two-epoch discretization with separate GY94 (Goldman and Yang 1994) codon substitution models.
- Parameters of interest: ω_1 (dN/dS before progression time T_1) and ω_2 (dN/dS after progression).
- HPD intervals show a general decrease in parameter estimates after progression time.

Within-host HIV evolutionary dynamics

Patient	log Bayes factor
patient 1	7.418
patient 2	9.602
patient 3	2.174
patient 5	-0.282
patient 6	9.21
patient 7	8.112
patient 8	2.627
patient 9	2.142
Joint evidence:	2.14

- Bayes factor tests for $I(\omega_2 < \omega_1)$ (dn/ds after progression time $< dn/ds$ before progression time).
- Strong evidence for patients 1, 2, 6 and 7.
- Joint Bayes factor test for posterior odds over the prior odds that $\omega_1 < \omega_2$ supports the immune relaxation hypothesis.

Within-host HIV evolutionary dynamics

Patient	log Bayes factor
patient 1	7.418
patient 2	9.602
patient 3	2.174
patient 5	-0.282
patient 6	9.21
patient 7	8.112
patient 8	2.627
patient 9	2.142
Joint evidence:	2.14

- Bayes factor tests for $I(\omega_2 < \omega_1)$ (dn/ds after progression time $< dn/ds$ before progression time).
- Strong evidence for patients 1, 2, 6 and 7.
- Joint Bayes factor test for posterior odds over the prior odds that $\omega_1 < \omega_2$ supports the immune relaxation hypothesis.

Seasonal circulation of Influenza A H3N2

- Influenza A H3N2 sequences sampled between 2003 to 2006 from Australia, Europe, Japan, USA, New Zealand, Southeast Asia and Hong Kong (Bahl *et al.* 2011).
- Data at the N tips of topology $\mathbf{F}$ consisting of character sequence data $\mathbf{X} = (X_1, ..., X_N)$ and spatial locations $\mathbf{Y} = (Y_1, ..., Y_N)$ is generated by independent stochastic processes (Lemey *et al.* 2009).
- $P(\mathbf{F}, Q, \phi | \mathbf{X}, \mathbf{Y}) \propto$

$$\underbrace{P(\mathbf{X}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{Y}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{F})}_{\substack{topology \\ prior}} \cdot \underbrace{P(Q)}_{\substack{substitution \\ prior}} \cdot \underbrace{P(\phi)}_{\substack{substitution \\ prior}}$$

Seasonal circulation of Influenza A H3N2

- Influenza A H3N2 sequences sampled between 2003 to 2006 from Australia, Europe, Japan, USA, New Zealand, Southeast Asia and Hong Kong (Bahl *et al.* 2011).
- Data at the N tips of topology $\mathbf{F}$ consisting of character sequence data $\mathbf{X} = (X_1, ..., X_N)$ and spatial locations $\mathbf{Y} = (Y_1, ..., Y_N)$ is generated by independent stochastic processes (Lemey *et al.* 2009).
- $$P(\mathbf{F}, Q, \phi | \mathbf{X}, \mathbf{Y}) \propto \underbrace{P(\mathbf{X}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{Y}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{F})}_{\substack{topology \\ prior}} \cdot \underbrace{P(Q)}_{\substack{substitution \\ prior}} \cdot \underbrace{P(\phi)}_{\substack{substitution \\ prior}}$$

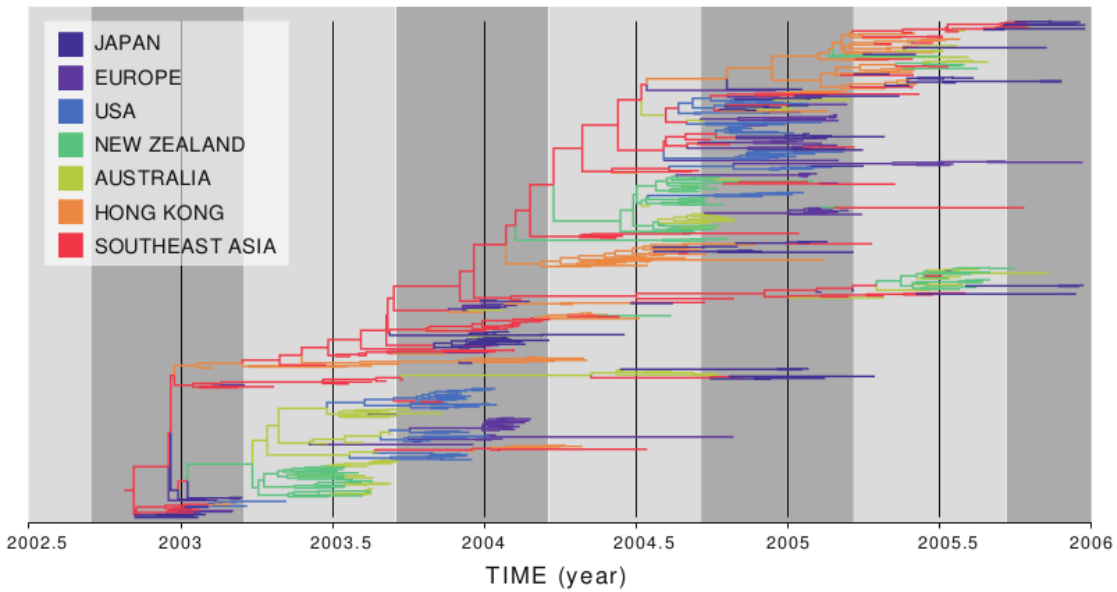
Seasonal circulation of Influenza A H3N2

- Influenza A H3N2 sequences sampled between 2003 to 2006 from Australia, Europe, Japan, USA, New Zealand, Southeast Asia and Hong Kong (Bahl *et al.* 2011).
- Data at the N tips of topology $\mathbf{F}$ consisting of character sequence data $\mathbf{X} = (X_1, ..., X_N)$ and spatial locations $\mathbf{Y} = (Y_1, ..., Y_N)$ is generated by independent stochastic processes (Lemey *et al.* 2009).
- $P(\mathbf{F}, Q, \phi | \mathbf{X}, \mathbf{Y}) \propto$

$$\underbrace{P(\mathbf{X}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{Y}|\mathbf{F})}_{Likelihood} \cdot \underbrace{P(\mathbf{F})}_{\substack{topology \\ prior}} \cdot \underbrace{P(Q)}_{\substack{substitution \\ prior}} \cdot \underbrace{P(\phi)}_{\substack{substitution \\ prior}}$$

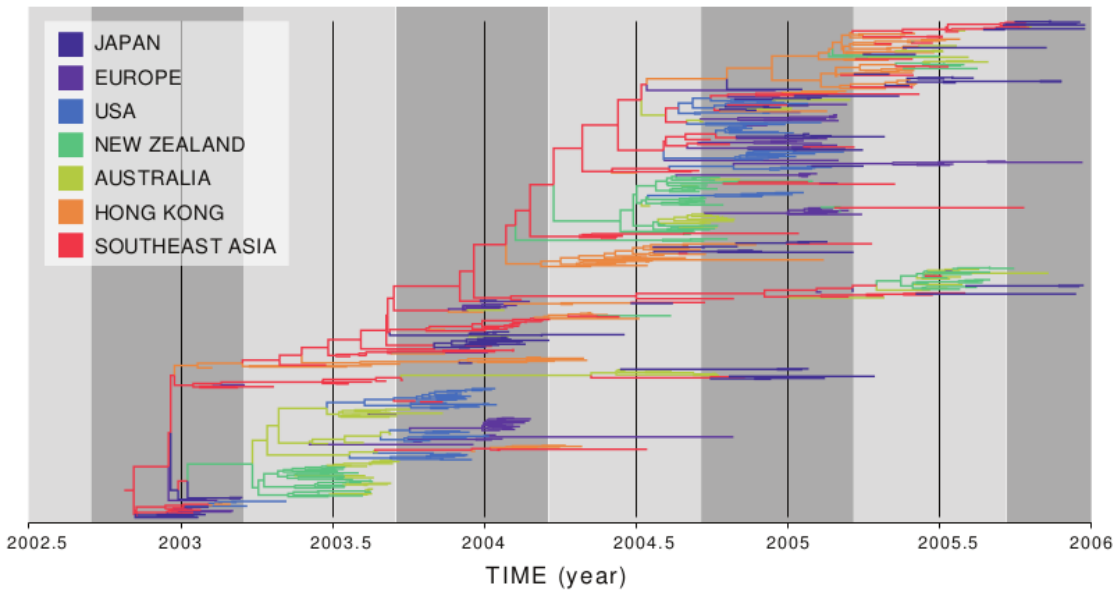
Seasonal circulation of flu

- We extend this discrete phylogeographic approach by specifying epochs alternating between spring and summer to autumn and winter on the northern hemisphere.
- Parameters of the discrete diffusion processes are shared among spring - summer as well as autumn - winter epochs, effectively creating two separate rate matrices.
- This is schematically represented in the figure.



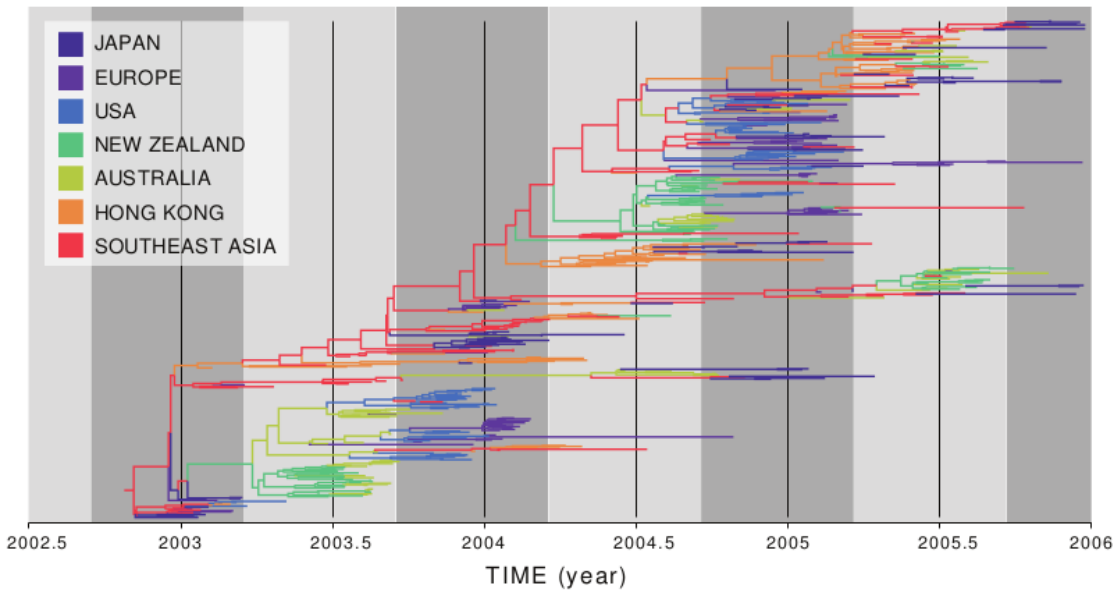
Seasonal circulation of flu

- We extend this discrete phylogeographic approach by specifying epochs alternating between spring and summer to autumn and winter on the northern hemisphere.
- Parameters of the discrete diffusion processes are shared among spring - summer as well as autumn - winter epochs, effectively creating two separate rate matrices.
- This is schematically represented in the figure.

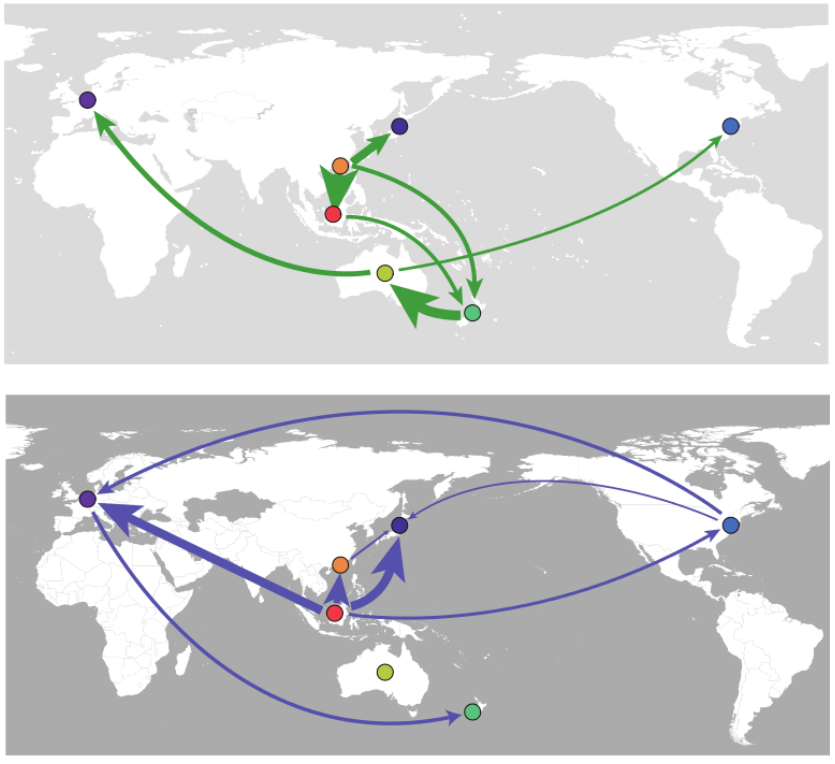


Seasonal circulation of flu

- We extend this discrete phylogeographic approach by specifying epochs alternating between spring and summer to autumn and winter on the northern hemisphere.
- Parameters of the discrete diffusion processes are shared among spring - summer as well as autumn - winter epochs, effectively creating two separate rate matrices.
- This is schematically represented in the figure.

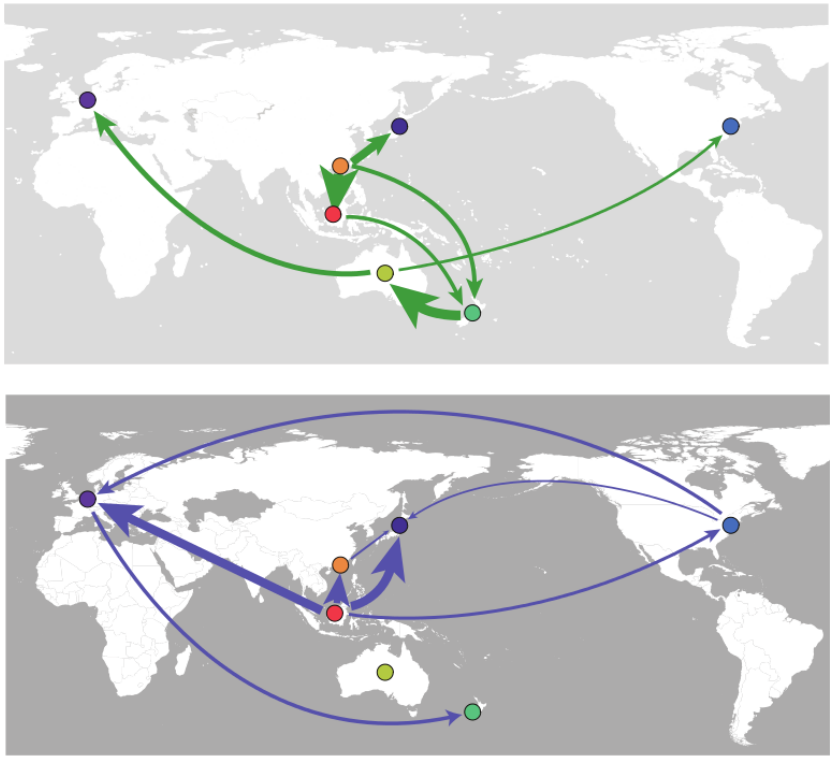


Seasonal circulation of flu



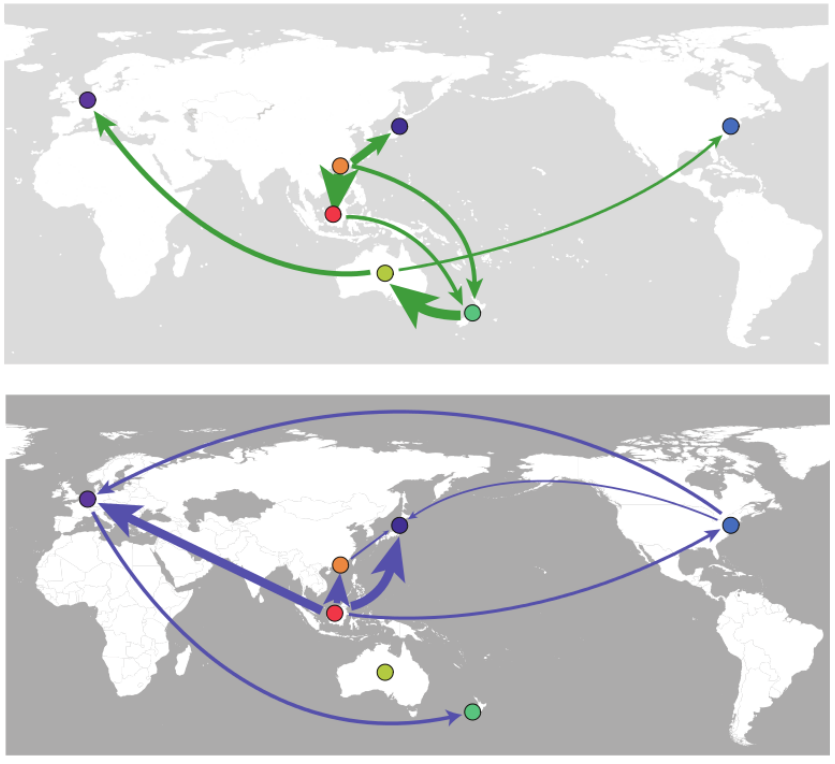
- Bayesian stochastic search variable selection (BSSVS, Lemey *et al.* 2009) procedure is used to identify best supported subset of diffusion rates within each epoch.
- Figure on the left presents rates yielding a Bayes factor > 20 for the spring - summer and autumn - winter epochs respectively.
- Findings suggest seasonal dynamics, with spring and summer diffusion mirroring autumn and winter.

Seasonal circulation of flu



- Bayesian stochastic search variable selection (BSSVS, Lemey *et al.* 2009) procedure is used to identify best supported subset of diffusion rates within each epoch.
- Figure on the left presents rates yielding a Bayes factor > 20 for the spring - summer and autumn - winter epochs respectively.
- Findings suggest seasonal dynamics, with spring and summer diffusion mirroring autumn and winter.

Seasonal circulation of flu



- Bayesian stochastic search variable selection (BSSVS, Lemey *et al.* 2009) procedure is used to identify best supported subset of diffusion rates within each epoch.
- Figure on the left presents rates yielding a Bayes factor > 20 for the spring - summer and autumn - winter epochs respectively.
- Findings suggest seasonal dynamics, with spring and summer diffusion mirroring autumn and winter.

Conclusions

- Evolutionary inference approach relaxing the standard time-homogeneity assumption.
- Applicable to any discrete data type.
- Implemented in BEAST/BEAGLE Bayesian phylogenetic framework.
- **TODO**: Estimate transition times, model parameter change as a continuous function of time, further stretch computational limits

Software availability

- BEAST (Bayesian Evolutionary Analysis by Sampling Trees) source code is freely available at http://code.google.com/p/beast-mcmc/ under the terms of GNU LGPL license.
- BEAGLE (Broad-platform Evolutionary Analysis General Likelihood Evaluator) library is free, open-source software licensed under the GNU LGPL. Both the source code and binary installers are available from www.code.google.com/p/beagle-lib/.
- SPREAD (Spatial Phylogenetic Reconstruction of Evolutionary Dynamics) visualisation software is licensed under the GNU Lesser GPL, and its source code is freely available at https://github.com/phylogeography/SPREAD. Compiled, runnable packages are hosted at http://www.phylogeography.org/SPREAD.html.

Thanks to...

European Research Council

- Philippe Lemey

  Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
- Guy Baele

  Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
- Andrew Rambaut

  Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
- Marc Suchard

  Department of Biomathematics, University of California, Los Angeles, USA