A statistical learning approach to infer transmissions of infectious diseases from deep sequencing data

Maryam Alamil, Joseph Hughes, Karine Berthier, Cecile Desbiez, Gaël Thébaud, Samuel Soubeyrand

A statistical learning approach to infer transmissions of infectious diseases from deep sequencing data

Maryam Alamil[1], Joseph Hughes[2], Karine Berthier[3], Cécile Desbiez[3], Gaël Thébaud[4] and Samuel Soubeyrand[1].

[1]BioSP, INRA, 84914, Avignon, France
[2]MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom
[3]Pathologie Végétale, INRA, 84140 Montfavet, France
[4]BGPI, INRA, SupAgro, Cirad, Univ. Montpellier, Montpellier, France

12 March 2019

Overview

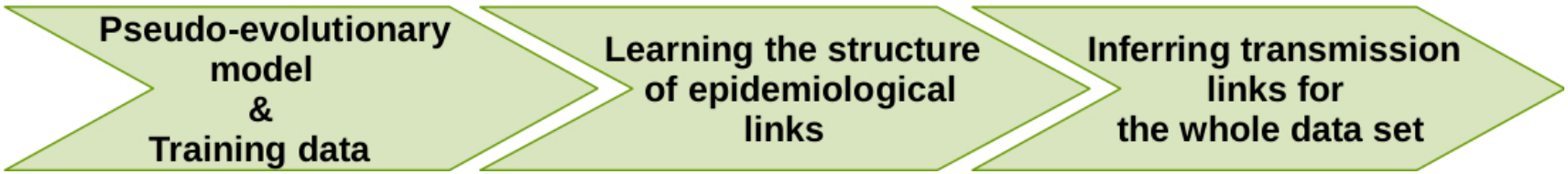
Overview

Introduction

- Inferring transmission links, for fast evolving pathogens, using viral genetic data is crucial to make epidemiological predictions and to design control strategies.
- Pathogen sequence data have been exploited to infer who infected whom, by using empirical and model-based approaches.
- Data collected with deep sequencing techniques provide a subsample of the pathogen variants that were present in the host at sampling time. They are expected to give better insight into epidemiological links.

Introduction

- Here, we present a Statistical Learning Approach For Estimating Epidemiological Links from deep sequencing data (SLAFEEL), which is summarized as follows:



- After that, we apply this approach to three real cases of animal, human and plant epidemics.

- Then, we show the impact of introducing penalization and, therefore, using training data on the inference.

Overview

Pseudo-evolutionary model for the evolution and transmission of populations of sequences

- ☞ It describes transitions between sets of sequences sampled at different times from an infected host and its putative sources.
- ☞ It takes into consideration the loss and gain of virus variants during within-host evolution and their loss during between-host transmissions.
- ☞ We built a sort of regression function parameterised by an evolutionary parameter and a penalisation parameter, where:
  - the response variable is the set of sequences $S = \{S_1, ..., S_J\}$ observed from a recipient host unit,
  - the explanatory variable is the set of sequences $S^{(0)} = \{S_1^{(0)}, ..., S_I^{(0)}\}$ observed from a putative source.
  - the coefficients are weights measuring how each sequence in $S^{(0)}$ contributes to explaining each sequence in S.

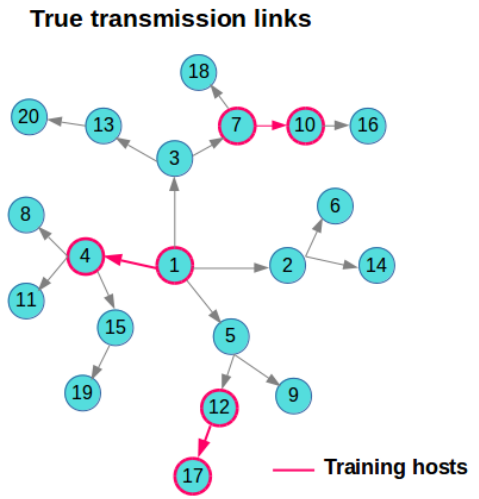
Estimation and calibration of parameters, and inference of transmissions

Step 1 Conditional estimation of the evolutionary parameter μ given θ

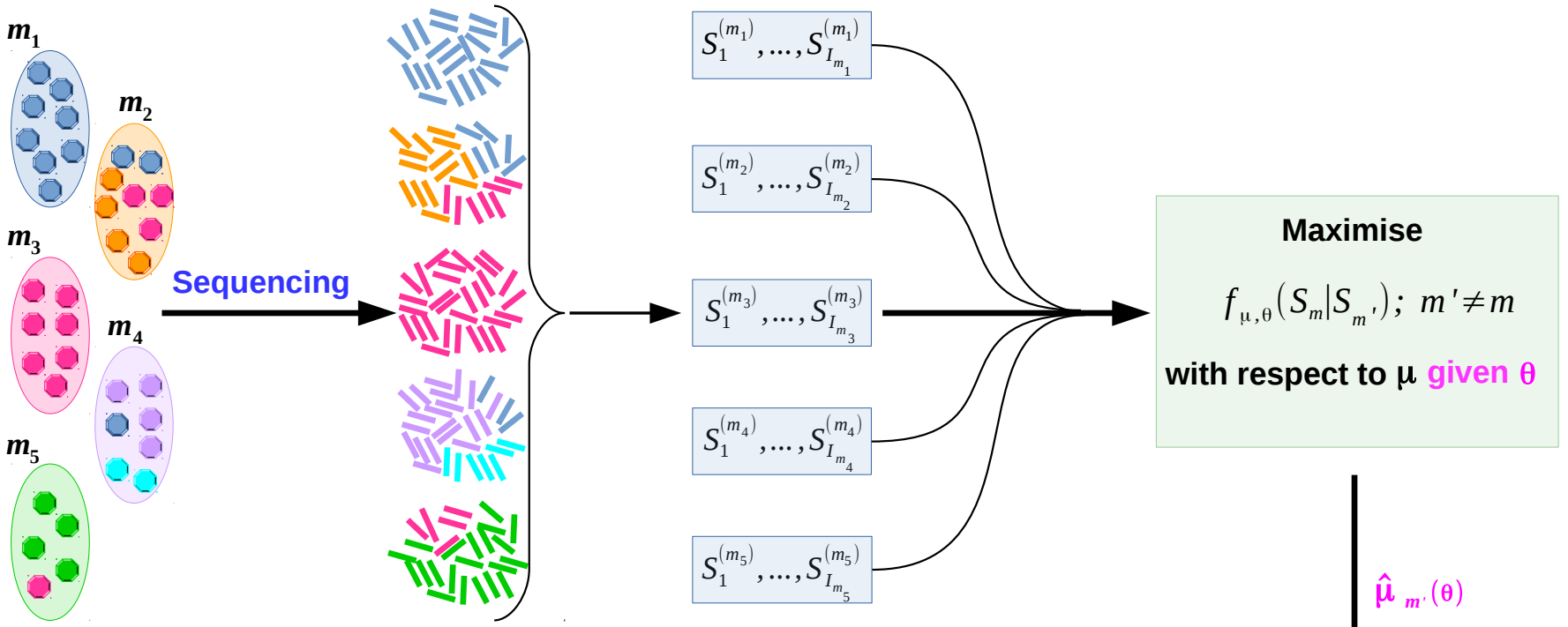
Step 2 Calibration of the penalization parameter θ
(Based on a statistical learning approach)

Step 3 Inference of transmission links

Training hosts



Estimation and calibration of parameters, and inference of transmissions



Semi-parametric pseudo-evolutionary model (1/2)

☞ Its general form is given by a penalized pseudo-likelihood:

$$f_{\mu,\theta}\left(S_1,...,S_J|S_1^{(0)},...,S_I^{(0)}\right) = P_\theta(W) \times \prod_{j=1}^{J} \left(\underbrace{\frac{\sum_{i=1}^{I} w_{ij} K_\mu\{d(S_j, S_i^{(0)}); \Delta_{ij}\}}{\sum_{i=1}^{I} w_{ij}}}_{p_j} \right)$$

where:

- $d(.,.)$ is a distance function giving the number of different nucleotides between two sequences,
- Δ_{ij} is the duration separating the two sequences S_j and $S_i^{(0)}$,
- $w_{ij} = 1/n_j$ for indices i corresponding to sequences $S_i^{(0)}$ minimally distant from the sequence S_j (the number of such sequences denote n_j) and $w_{ij} = 0$ otherwise,
- $K_\mu(., \Delta)$ is a kernel smoother parameterised by an evolutionary parameter μ. It is the probability distribution function of the binomial law with size equals to the sequence length and success probability $3(1 - exp(-4\mu\Delta))/4$,

Semi-parametric pseudo-evolutionary model (2/2)

- $P_\theta(W)$ is a parametric penalization for the weight matrix W, parameterised by a penalisation parameter θ. It measures the likelihood of the contributions of explanatory sequences $S_1^{(0)}, ..., S_I^{(0)}$ (measured by $\sum_{j=1}^{J} w_{ij}$, $i = 1, ..., I$) to the response set of sequences $S_1, ..., S_J$.

- Two hypotheses are considered for the penalization:

  1. $H_1 : \mathbb{E}\left[\sum_{j=1}^{J} w_{ij}\right] = J/I$. Two associated penalization shapes:

     - $P_\theta(W) = \prod_{i=1}^{I} \Phi\left(\sum_{j=1}^{J} w_{ij}; \frac{J}{I}, \theta\frac{J}{I}\left(1 - \frac{1}{I}\right)\right)$,

     - $P_\theta(W) = \theta\chi^2\left(\sum_{i=1}^{I} \frac{\left(\sum_{j=1}^{J} w_{ij} - J/I\right)^2}{J/I}; I - 1\right)$,

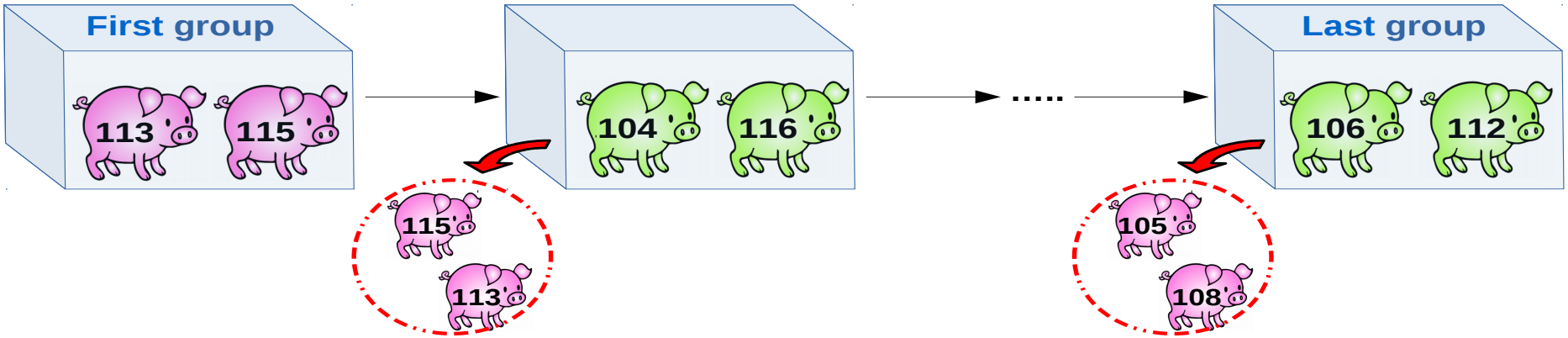
  2. $H_2 : \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} min(d(S_j, S_{f(j)}^{(0)}))\right] = \bar{d}_{obs}$. A linked penalization shape:

     - $P_\theta(W) = \theta\prod_{j=1}^{J} \Phi\left(\sum_{i=1}^{I} w_{ij} d(S_j, S_i^{(0)}); \bar{d}_{\text{obs}}, \sigma^2_{\text{obs}}\right)$.

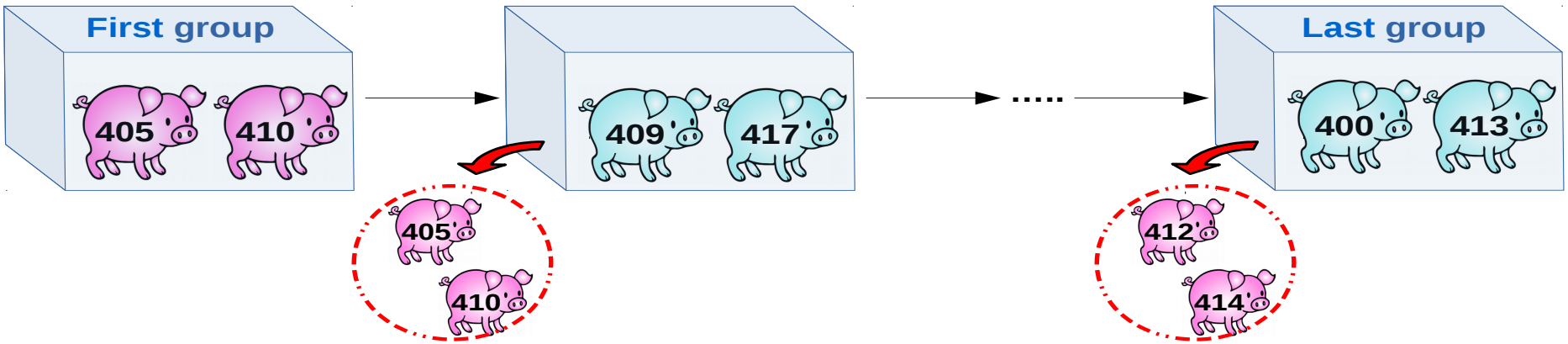
Overview

Inference of epidemiological links for hosts with different immunological histories

Naive chain (5 groups)



Vaccinated chain (7 groups)



Inference of epidemiological links for hosts with different immunological histories

Naive chain of Influenza	
Hosts	**Contact**
113	115
115	113
104	113, 115, 116
116	104, 113, 115
109	104, 111, 116
111	104, 109, 116
105	108, 109, 111
108	105, 109, 111
106	105, 108, 112
112	105, 106, 108

Vaccinated chain of Influenza	
Hosts	**Contact**
405	410
410	405
409	405, 410, 417
417	405, 409, 410
401	409, 417
415	401, 416
416	401, 415
403	406, 415, 416
406	403, 415, 416
412	403, 406, 414
414	403, 406, 412
400	412, 413, 414
413	400, 412, 414

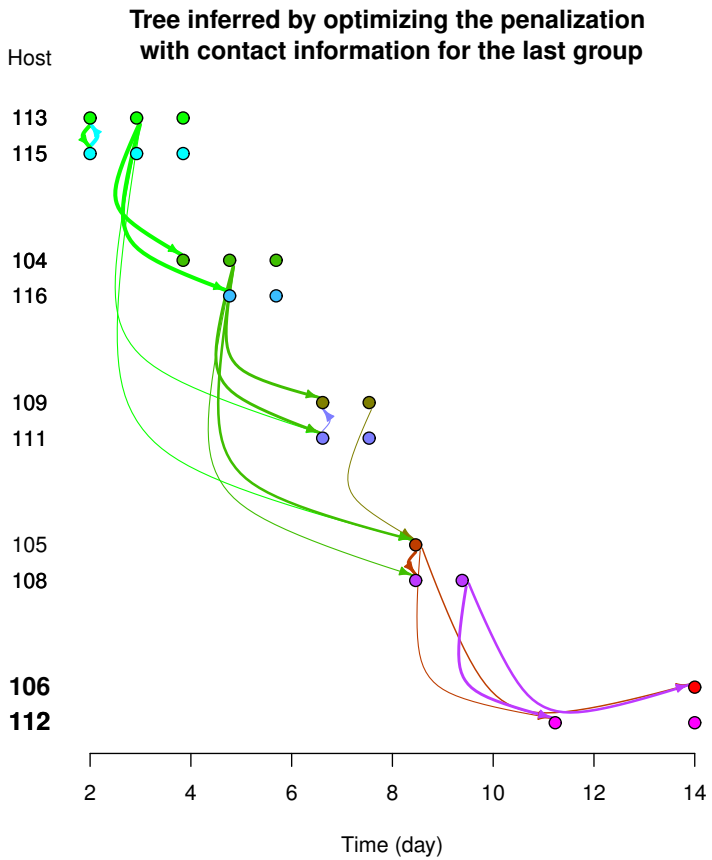
Inference of epidemiological links for hosts with different immunological histories

Naive chain of Influenza	
Hosts	**Contact**
113	115
115	113
104	113, 115, 116
116	104, 113, 115
109	104, 111, 116
111	104, 109, 116
105	108, 109, 111
108	105, 109, 111
106	105, 108, 112
112	105, 106, 108

Vaccinated chain of Influenza	
Hosts	**Contact**
405	410
410	405
409	405, 410, 417
417	405, 409, 410
401	409, 417
415	401, 416
416	401, 415
403	406, 415, 416
406	403, 415, 416
412	403, 406, 414
414	403, 406, 412
400	412, 413, 414
413	400, 412, 414

Inference of epidemiological links for hosts with different immunological histories

A



B

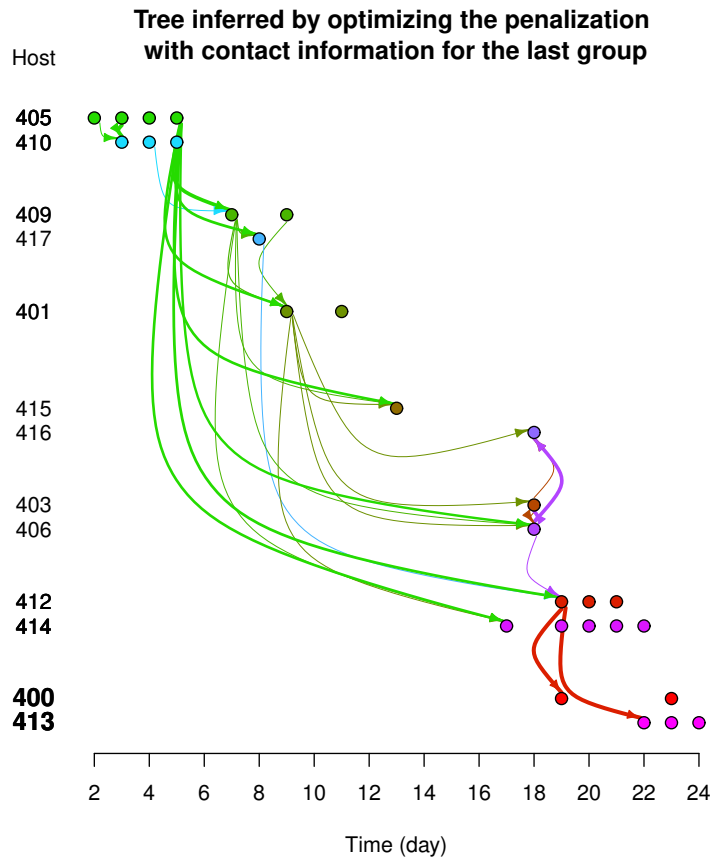
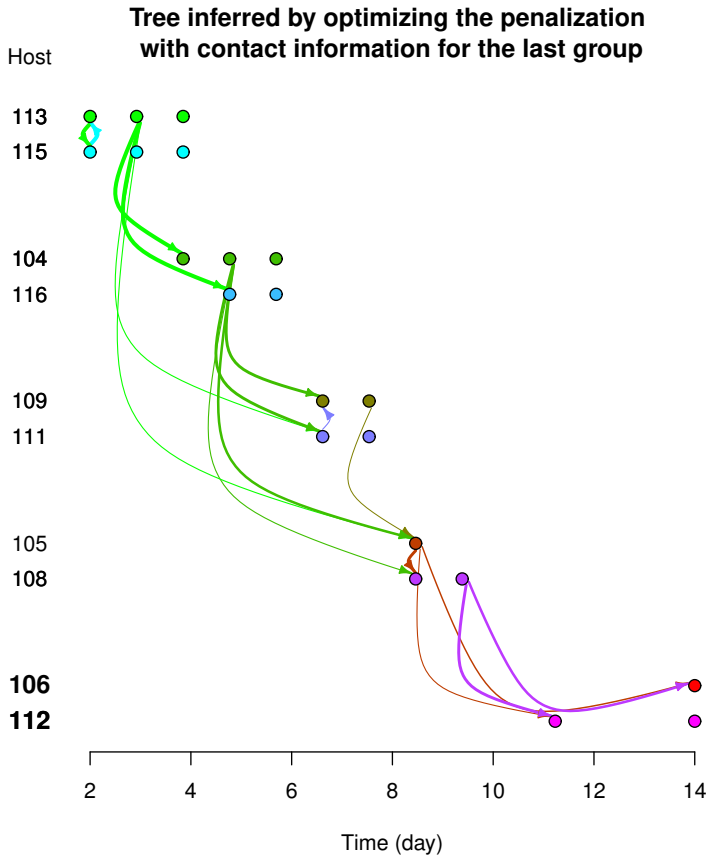


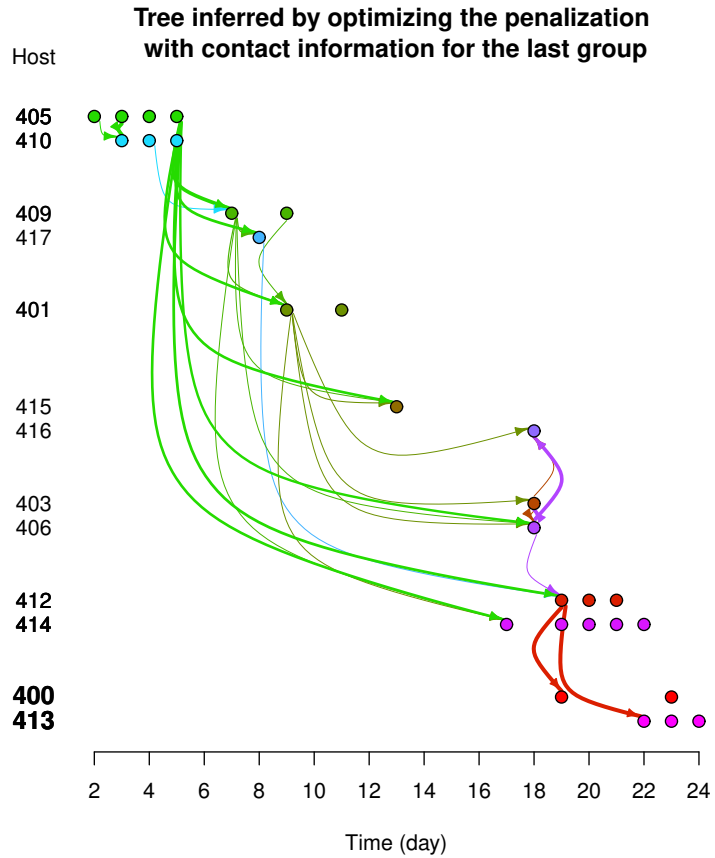
Figure: Transmissions inferred in the naive chain (A) and vaccinated chain (B) of Swine influenza virus using pair of training hosts in the last group for calibrating the penalization. Training hosts are written in bold. The thickness of each arrow is proportional to the intensity of the corresponding inferred link.

Inference of epidemiological links for hosts with different immunological histories

A



B



Consistent estimations with the two pairs of training hosts.

Impact of training hosts

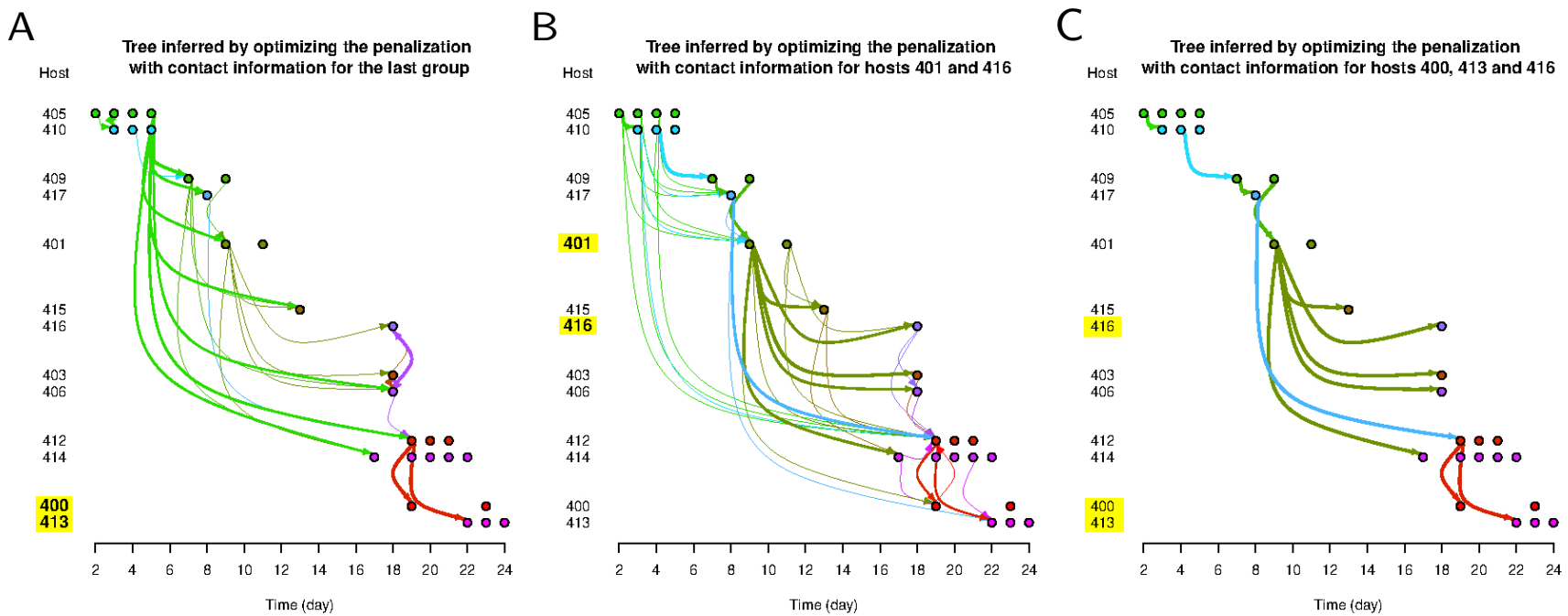
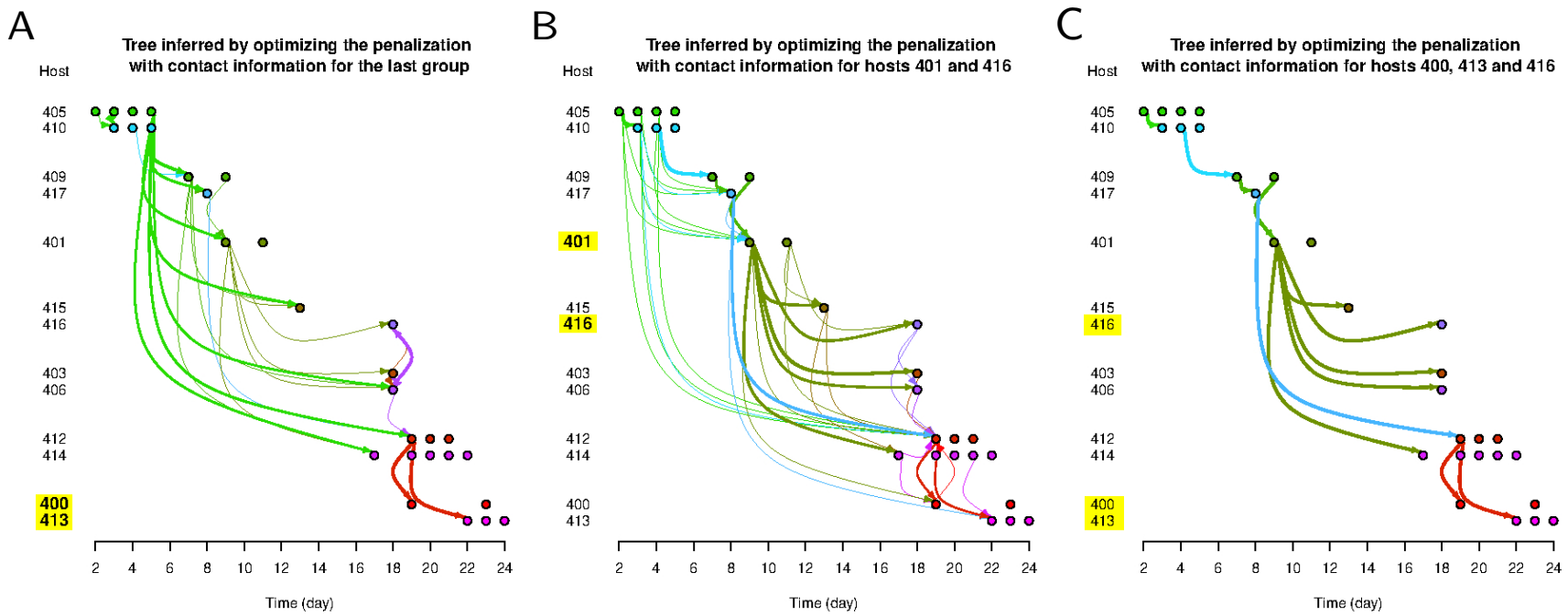


Figure: Transmissions inferred in the vaccinated chain of SIV using different sets of training hosts for calibrating the penalization: (A) a pair of hosts in the last group of the chain (B) a pair of hosts in the middle of the chain (C) three hosts in the last group and the middle of the chain. Training hosts are highlighted. The thickness of each arrow is proportional to the intensity of the corresponding inferred link.

Impact of training hosts



Using more contact information allows a finer calibration of the penalisation and, consequently, a more accurate resolution of transmissions.

Comparaison between SLAFEEL and BadTrIP

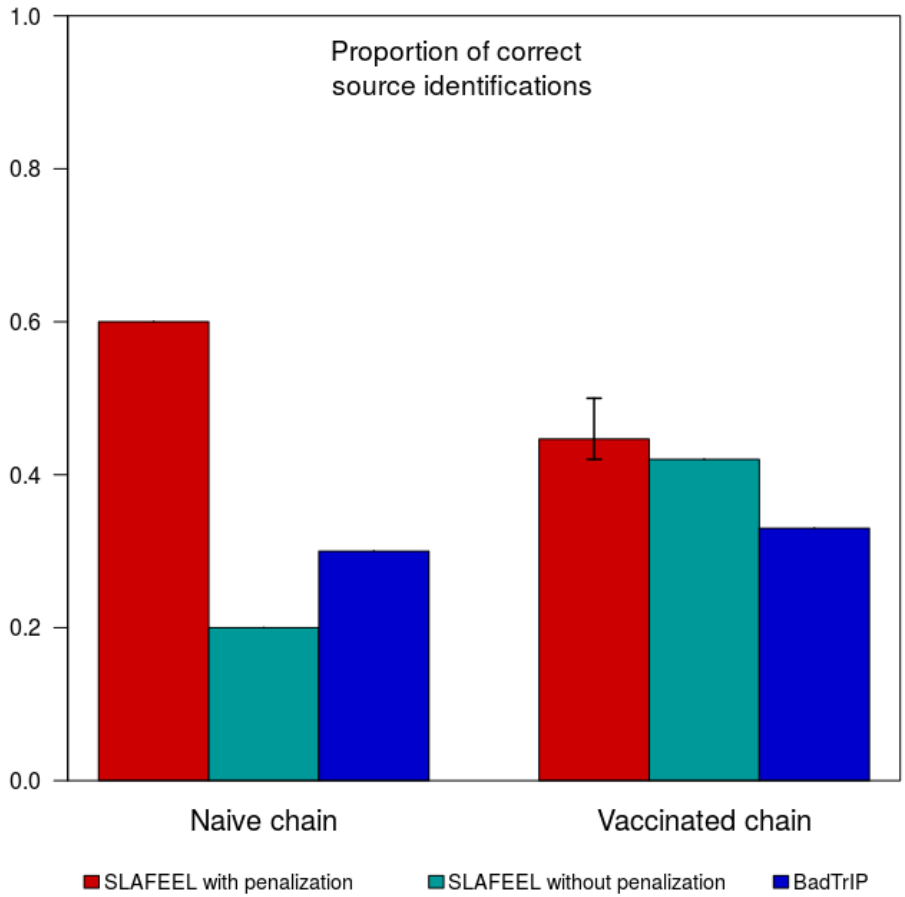


Figure: Discrepancy between inferred transmission graphs obtained with SLAFFEL (with and without penalization) and BadTrIP and reference graphs, for naive and vaccinated chains of SIV. This discrepancy is measured by the proportion of correct source identifications.

Inference of epidemiological links in a low diversity pathogen population

Recipient	Donor
G3817	G3729
G3820	G3729
G3821	G3729
G3823	G3729
G3851	G3752

Senga et al. (2017)

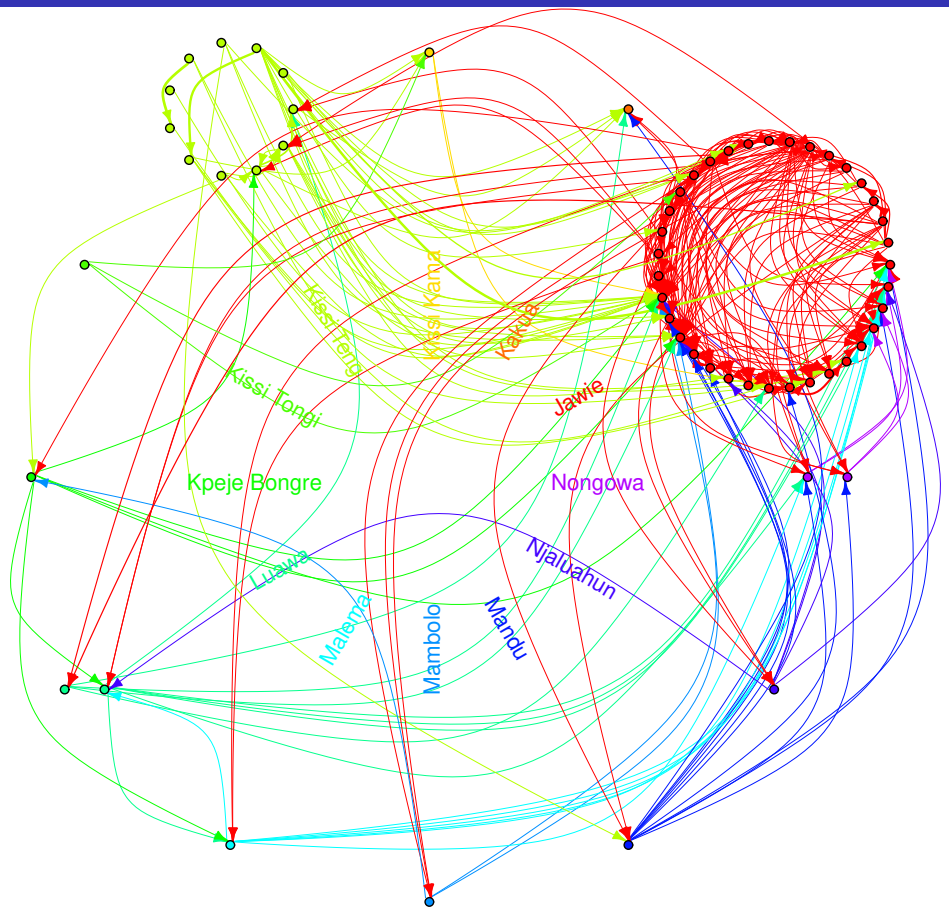
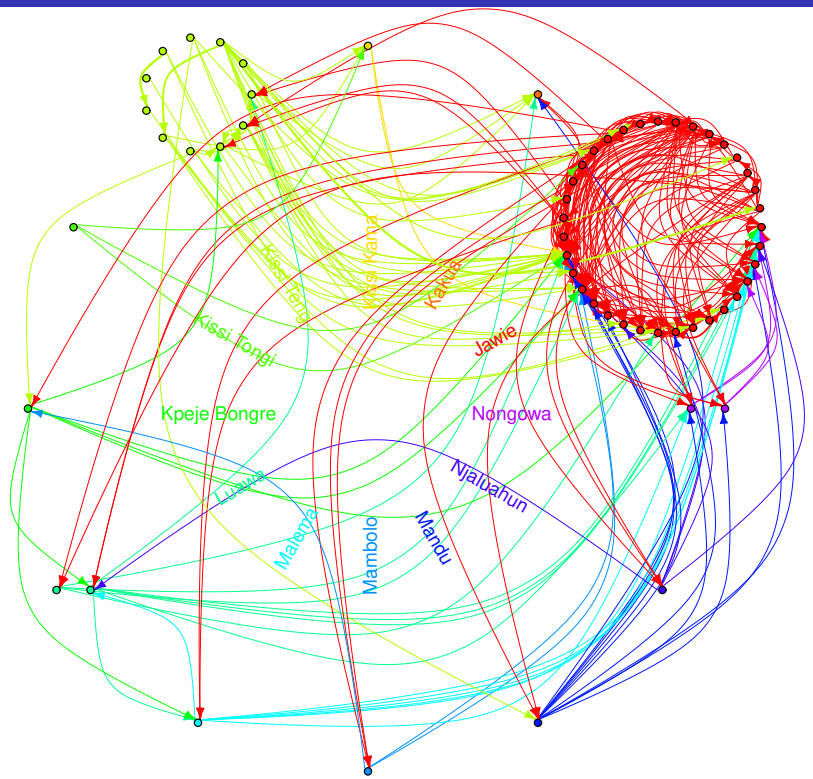


Figure: Most likely epidemiological links between Ebola patients cumulating to 20% probability for each recipient (i.e., for each recipient, potential donors were ranked with respect to link intensity, and the subset of donors with higher ranks for which the sum of link intensities reached 0.2 were retained to be displayed in the graph).

Inference of epidemiological links in a low diversity pathogen population

Recipient	Donor
G3817	G3729
G3820	G3729
G3821	G3729
G3823	G3729
G3851	G3752

Senga et al. (2017)



The Jawie chiefdom seems to be an interface between Kissi Teng and Kissi Tongi chiefdoms on the one hand and most of the other chiefdoms on theother hand.

Inference of epidemiological links at a metapopulation scale

☞ Geographic proximity is used as contact information

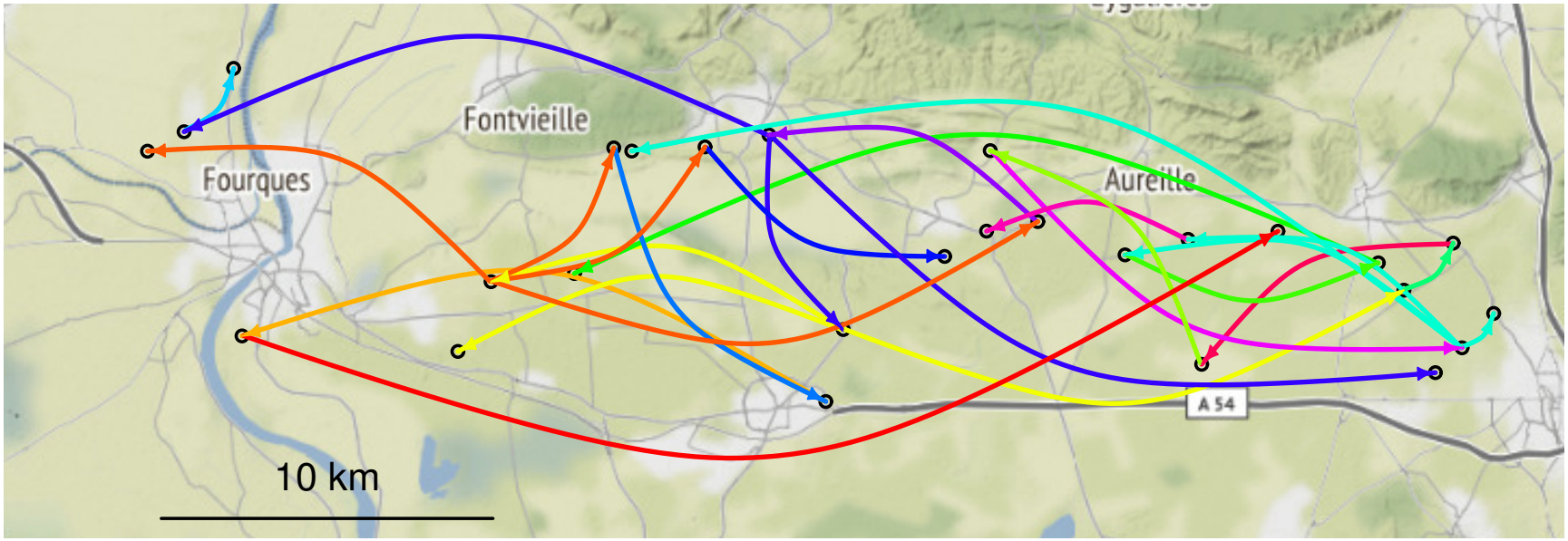
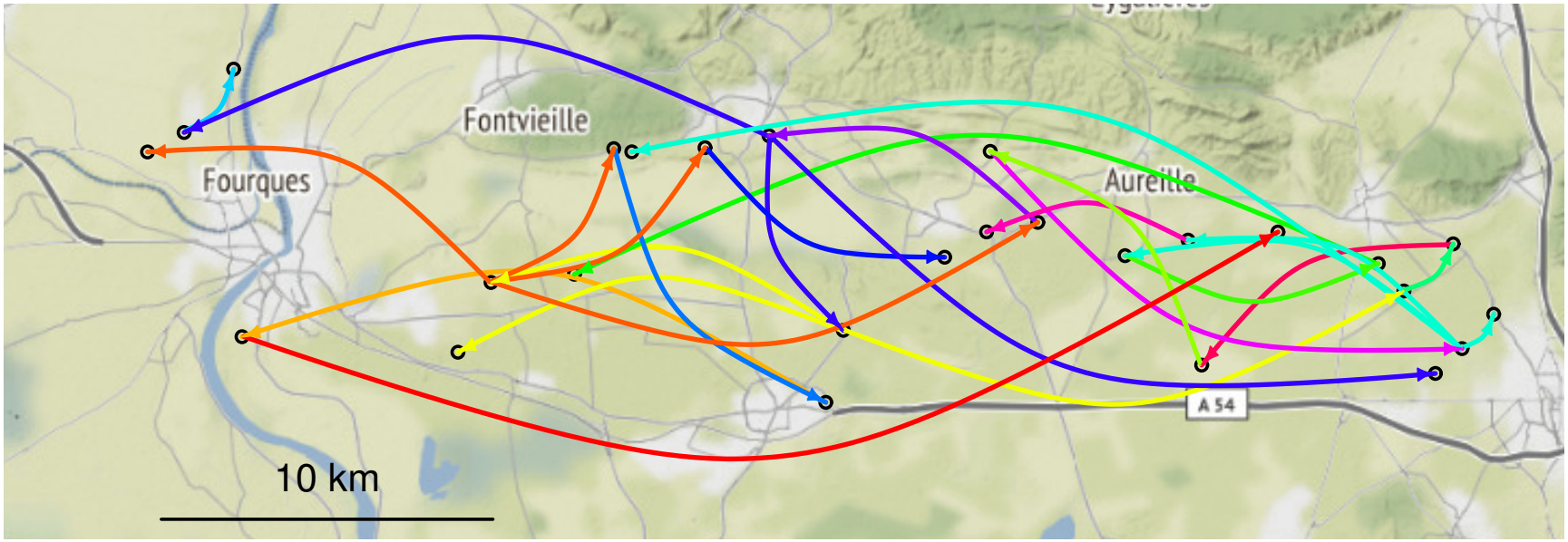


Figure: Epidemiological links inferred between 27 salsify patches based on sets of potyvirus variants sequenced from 189 infected plants sampled in a 40 × 10 km region of south-eastern France.

Inference of epidemiological links at a metapopulation scale



- **No secondary arrows are displayed,**
- **Non-negligible proportion of long links,**
- **Environmental factors and intra-host demo-genetic factors may play a role in the transmission of the virus.**

Overview

Conclusion and perspectives

- SLAFEEL is adaptable to very different contexts and data from animal, human and plant epidemics.
- SLAFEEL is valuable in non-standard situations where classical mechanistic assumptions may be erroneous and when sequencing and variant calling may be noisy.
- Introducing a penalization and using more contact information lead to accurate inferences of transmission links.
- Calibrate and assess its efficiency by applying it to simulated data generated with diverse sampling efforts, sequencing techniques and stochastic models of viral evolution and transmission.
- Investigate the statistical relationship between inferred transmission links and environmental factors.

References

M Alamil, J Hughes, K Berthier, C Desbiez, G Thébaud, and S Soubeyrand.

Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases.

Submitted.

C Desbiez, A Schoeny, B Maisonneuve, K Berthier, et al.

Molecular and biological characterization of two potyviruses infecting lettuce in southeastern france.

Plant Pathology, 66(6):970–979, 2017.

SK Gire, A Goba, KG Andersen, RS Sealfon, et al.

Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.

Science, 345:1369–1372, 2014.

PR Murcia, J Hughes, P Battista, L Lloyd, GJ Baillie, et al.

Evolution of an Eurasian Avian-like influenza virus in naïve and vaccinated pigs.

PLoS Pathogens, 8(5):e1002730, 2012.

Thank you for your attention!

We welcome your questions, comments & suggestions!

Impact of introducing a penalization

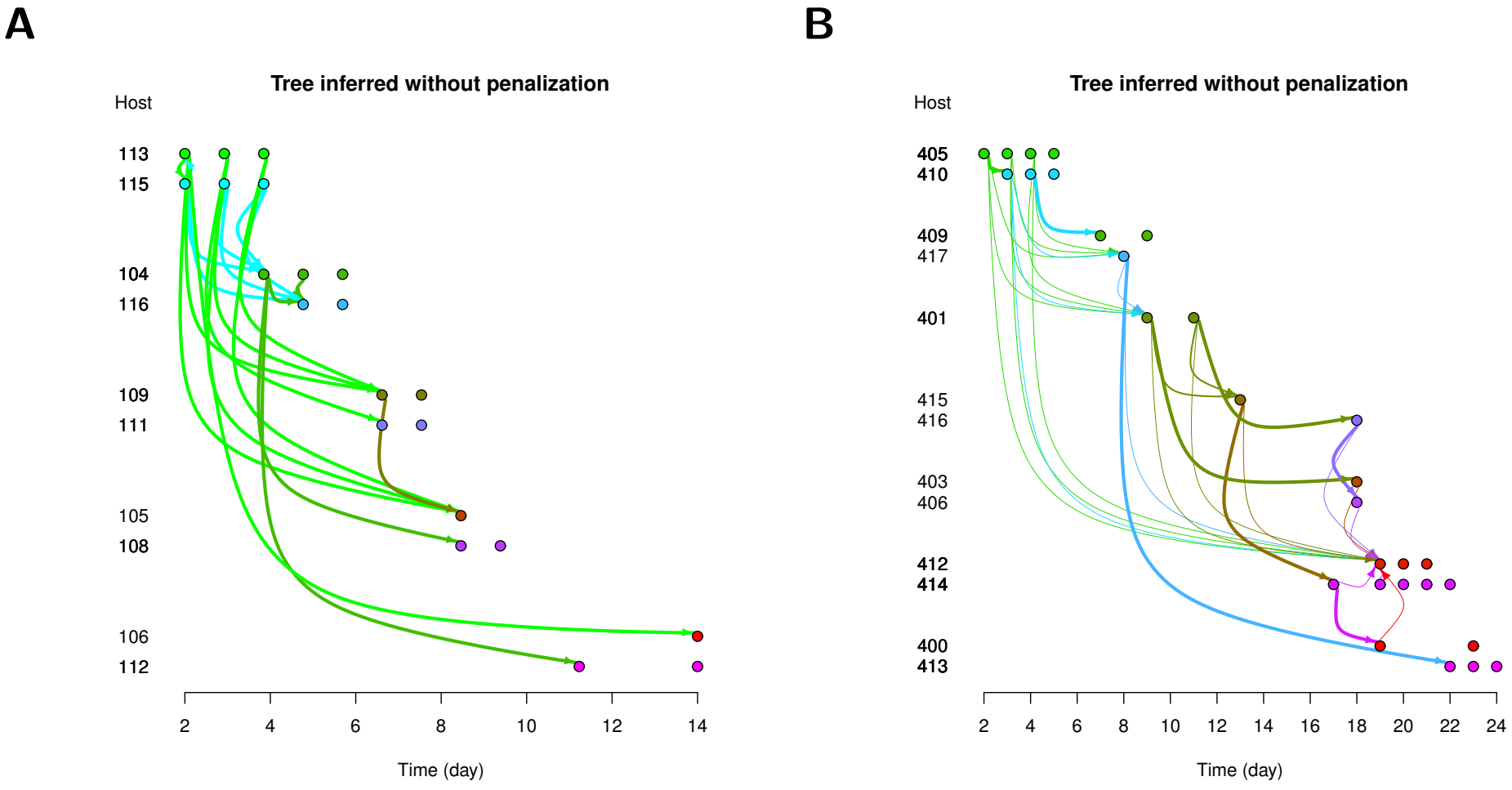


Figure: Transmissions inferred in the naive chain (A) and vaccinated chain (B) of SIV without including the penalisation and, therefore, without including training hosts.