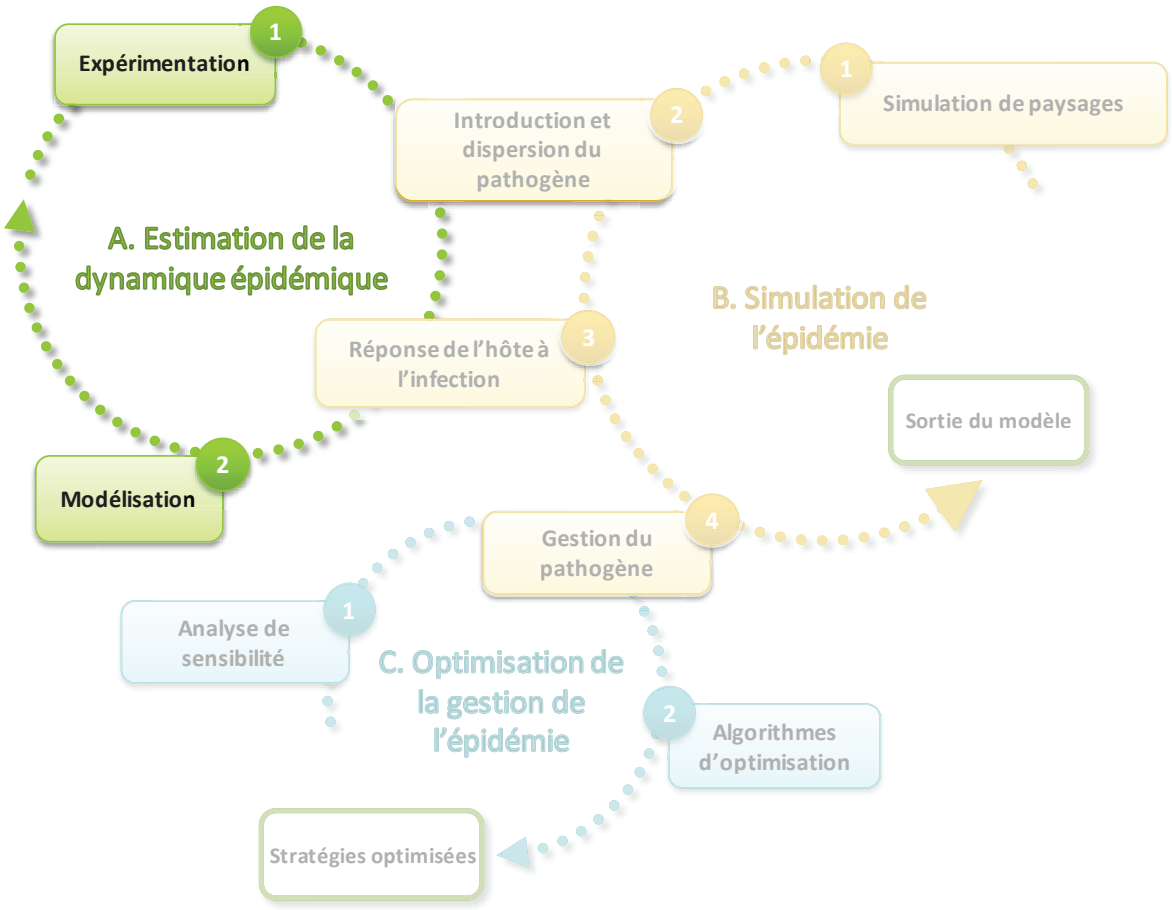
ESTIMATION DE PARAMETRES EPIDEMIOLOGIQUES



La première étape du processus de modélisation PESO consiste à estimer les paramètres qui caractérisent une épidémie. Comme cela a été présenté en introduction, les paramètres régissant les épidémies de sharka ont été estimés par Pleydell et al. (2018). Néanmoins, les données utilisées pour réaliser ces estimations ne prennent pas en compte la localisation exacte des arbres infectés, mais la proportion d'arbres infectés par parcelle, ce qui peut réduire la précision de l'estimation de la fonction de dispersion du virus. En effet, la connectivité des parcelles a été calculée à partir de leurs centroïdes, ce qui peut par exemple entrainer un biais dans l'estimation de la fonction de dispersion si un seul côté d'une parcelle comprend des arbres infectés. L'objectif de ce volet de ma thèse consiste à estimer de manière plus précise les paramètres épidémiologiques qui caractérisent la sharka, en utilisant des données épidémiologiques acquises au grain de l'arbre ainsi que des données génétiques.

Pour cela, j'ai tout d'abord réalisé une synthèse bibliographique (présentée dans la première partie de ce chapitre sous forme de revue) qui explique comment des données épidémiques et génétiques peuvent aider à la compréhension des épidémies. Dans une deuxième partie, nous avons tenté d'estimer plusieurs paramètres épidémiologiques de la sharka à l'aide d'un modèle visant à reconstruire les liens de transmission entre les hôtes individuels (en inférant « qui a infecté qui » dans le paysage).

1. Les modèles pour comprendre la dynamique des épidémies

Afin de comprendre la dynamique des épidémies, il est crucial d'identifier comment (voie de transmission), quand (période de transmission et fréquence), et où (hôte, emplacement et distance) ces pathogènes sont transmis. Pour cela, l'épidémiologie moléculaire est de plus en plus utilisée : cette approche exploite l'information sur la variabilité génétique des agents pathogènes pour caractériser leur dispersion et leur évolution. En particulier, des approches permettant d'estimer les paramètres épidémiologiques d'une maladie et d'identifier les voies de transmission de l'agent pathogène responsable entre les hôtes ou les populations hôtes ont été développées depuis une dizaine d'années.

La revue suivante présente certaines de ces approches qui exploitent l'information génétique pour suivre la dispersion d'un virus à travers un paysage. Dans le cadre de ma thèse, j'ai notamment contribué à l'écriture de la 3ème partie qui traite de l'inférence des arbres de transmission des maladies et de l'estimation des paramètres épidémiologiques, ainsi qu'à l'introduction et à la discussion.

ARTICLE 2

Exploiting Genetic Information to Trace Plant Virus Dispersal in Landscapes

Coralie Picard, Sylvie Dallot, Kirstyn Brunker, Karine Berthier, Philippe Roumagnac, Samuel Soubeyrand, Emmanuel Jacquot and Gaël Thébaud

Annual Review of Phytopathology

2017, Volume 55, Pages 139-160.

https://www.annualreviews.org/doi/full/10.1146/annurev-phyto-080516-035616

ANNUAL REVIEWS

Annual Review of Phytopathology

Exploiting Genetic Information to Trace Plant Virus Dispersal in Landscapes

Coralie Picard,[1,*] Sylvie Dallot,[1,*] Kirstyn Brunker,[2] Karine Berthier,[3] Philippe Roumagnac,[1] Samuel Soubeyrand,[4] Emmanuel Jacquot,[1] and Gaël Thébaud[1]

[1]UMR BGPI, INRA, Montpellier SupAgro, CIRAD, 34398, Montpellier Cedex 5, France; email: gael.thebaud@inra.fr

[2]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, United Kingdom

[3]Pathologie Végétale, INRA, 84140, Montfavet, France

[4]BioSP, INRA, 84914, Avignon, France

Annu. Rev. Phytopathol. 2017. 55:139–60

First published as a Review in Advance on May 19, 2017

The *Annual Review of Phytopathology* is online at phyto.annualreviews.org

https://doi.org/10.1146/annurev-phyto-080516-035616

*These authors contributed equally to this review.

Keywords

high-throughput sequencing, host range, invasion pathway, outbreak, transmission tree, vector

Abstract

During the past decade, knowledge of pathogen life history has greatly benefited from the advent and development of molecular epidemiology. This branch of epidemiology uses information on pathogen variation at the molecular level to gain insights into a pathogen's niche and evolution and to characterize pathogen dispersal within and between host populations. Here, we review molecular epidemiology approaches that have been developed to trace plant virus dispersal in landscapes. In particular, we highlight how virus molecular epidemiology, nourished with powerful sequencing technologies, can provide novel insights at the crossroads between the blooming fields of landscape genetics, phylogeography, and evolutionary epidemiology. We present existing approaches and their limitations and contributions to the

INTRODUCTION

bstitution rate: e of fixation of netic changes in a cies

Epidemics caused by the spread of pathogenic agents through host populations can be a high socioeconomic burden (71, 143). In order to support public policy decision-making regarding disease control strategies, scientists need to understand and, ultimately, quantify and predict how pathogens spread within and between host populations. This understanding has been recently improved by attempts to trace pathogen dispersal using molecular epidemiology and novel statistical approaches. Molecular epidemiology uses information on pathogen genetic variation to unravel the niche of a pathogen (including host and vector species) and characterize its dispersal and evolution (129). Such studies focus on the identification of risk factors that affect host exposure or intrinsic susceptibility to pathogens and on the dispersal of these pathogens from infected to susceptible hosts (8). In order to understand and control epidemics, it is indeed crucial to identify how (transmission route), when (transmission period and frequency), and where (host, location, and distance) pathogens are transmitted.

Although, ideally, fully documented epidemiological records would provide a wealth of necessary information, such a detailed level of pathogen-tracing information is not attainable in practice. However, even incomplete and indirect information on pathogen dispersal—such as host range, population connections, and epidemic origin and spread—can be highly valuable. In particular, the quantification of pathogen transmission across various distances, and specifically the characterization of long-distance dispersal events, has major implications for disease management strategies. To address these issues, pathogen tracing relies on indirect approaches that derive epidemiological information from the spatiotemporal structure of pathogen genetic diversity. Viruses are particularly amenable to such studies because their epidemiological and evolutionary dynamics occur at similar short timescales. Moreover, the high number of polymorphisms in their small genomes can be accessed relatively easily, and increasingly in real time, during epidemics (32, 60). As such, viruses are "measurably evolving" pathogens (7, 29).

lemental Material

The number of research articles published on virus molecular epidemiology has increased steadily since the 1990s—and since the 2000s for plant viruses (see **Supplemental Figure 1**). There are a few review articles on the use of plant virus diversity in evolutionary epidemiology (62, 92) or disease emergence studies (36, 40, 60, 105). As a complementary perspective, our purpose here is to specifically review molecular epidemiology approaches for plant viruses and to focus on how the molecular analysis of virus diversity provides insights into the spatiotemporal dynamics of plant virus epidemics. To this end, we explore three questions addressed by scientists in order to trace plant virus dispersal in landscapes: How to find the hosts and access virus diversity? What are the spatiotemporal history and predictors of virus flows in landscapes? How did the virus spread within an outbreak?

HOW TO ACCESS PATHOGEN DIVERSITY IN LANDSCAPES?

The vast majority of plant viruses have single-stranded DNA (ssDNA) or positive-stranded RNA genomes, which have a higher substitution rate (mostly ranging from 10^{-3} to 10^{-5} substitutions/site/year) than other genomes (7, 46, 125, 126). Proofreading-deficient polymerases, short generation times, and frequent bottlenecks on large populations all contribute to the impact of evolutionary forces on virus populations (38, 93) (see sidebar titled Evolutionary Processes Imprint Virus Genomes). Consequently, viral populations often show a high level of genetic diversity both within and between hosts (8, 32) (**Figure 1**). Characterizing the genetic structure and diversity

EVOLUTIONARY PROCESSES IMPRINT VIRUS GENOMES

Five evolutionary forces shape the genomes and genetic diversity of virus populations (60). The resulting patterns provide information on the underlying processes (50):

Mutation: The amount of de novo nucleotide diversity accessible across the spatiotemporal scales impacts the questions that molecular epidemiology approaches can address.

Recombination/reassortment: The generation of novel genetic combinations increases genomic diversity and thus adaptation. Viruses are mostly haploid and clonal, so this process is too infrequent to assume independence between loci.

Migration: The spatial reallocation of genomes increases genetic diversity within (and reduces differentiation between) populations. Estimation of migration rates provides key information on virus flows.

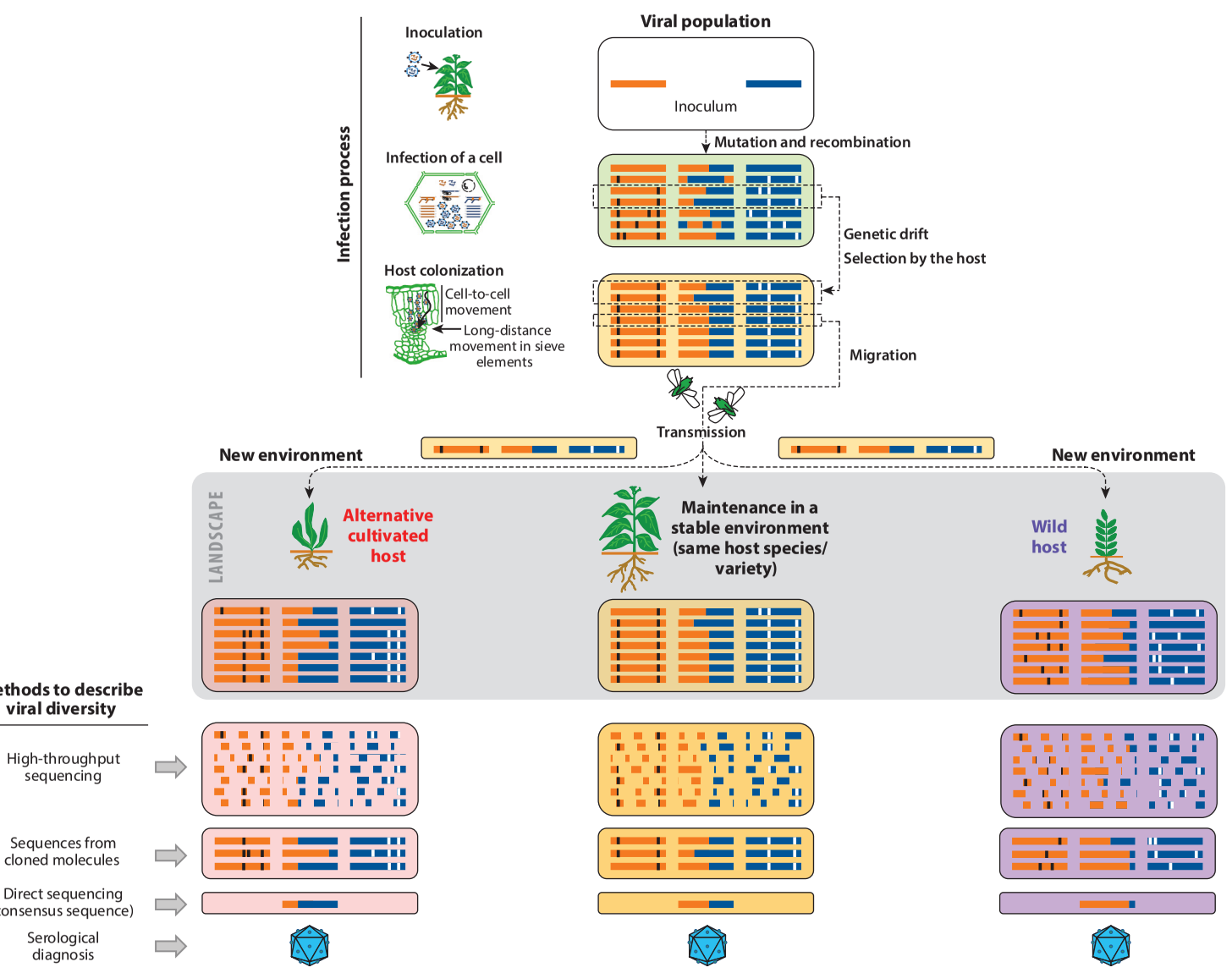
Selection: Fueled by the previous processes, selection is the engine of adaptation and can be stabilizing, directional, or diversifying. Environment-specific selection at some loci can pinpoint the original environment of a genome; however, highly reproducible mutational pathways toward adapted genotypes generate genetic homoplasy (i.e., shared polymorphism absent from the common ancestor) that can be misinterpreted as recombination and thus blur analyses.

Drift: The random sampling of individual genomes founding the next generation changes allele frequencies within a population. Drift promotes fixation of neutral (or slightly deleterious) mutations and thus increases differentiation between populations.

(*b*) the characteristics of the targeted potential hosts (plants/vectors and wild/cultivated and annual/perennial plants), (*c*) the spatiotemporal dynamics of the viral disease, and (*d*) the evolutionary rate of the virus under study. The past four decades have seen a huge evolution in the techniques used to reveal molecular polymorphisms and to sequence genomes.

Characterization of Virus Diversity

In order to propose a classification of viral species and to explore the diversity within virus species, the scientific community initially used biological properties of plant viruses, such as their host range, induced symptoms, and transmission properties, including the range of vectors involved. However, it was later shown that biological approaches are rarely adequate to reveal the structure and diversity of plant virus populations, as most of the polymorphisms of viral genomes have no effect on these biological parameters. In the 1970s, the development of techniques based on the antigenic properties of the capsid protein (18) shed a new light on the variability between and within viral species. Molecular techniques developed in the 1970s–1980s and widely used since the 1990s strongly modified plant virus epidemiology approaches. They allowed the direct characterization of viral genomes through the development of various molecular markers [e.g., restriction fragment length polymorphism (RFLP) (48), single-strand conformation polymorphism (SSCP) (102), ribonuclease protection assay (RPA) (42), and RNase T1 fingerprint (119)] and partial- or whole-genome sequencing using Sanger technology (124) on amplified [e.g., polymerase chain reaction (PCR) or rolling circle amplification (RCA)] products (128) or cloned molecules. Besides providing a quantitative estimate of the viral genetic diversity from within-host to global


Infection process
Inoculation
Infection of a cell
Host colonization
Cell-to-cell movement
Long-distance movement in sieve elements
Viral population
Inoculum
Mutation and recombination
Genetic drift
Selection by the host
Migration
Transmission
New environment
New environment
LANDSCAPE
Alternative cultivated host
Maintenance in a stable environment (same host species/ variety)
Wild host
ethods to describe viral diversity
High-throughput sequencing
Sequences from cloned molecules
Direct sequencing consensus sequence)
Serological diagnosis

Supplemental Material

RECOMBINANT GENOMES: TROUBLE OR TREASURE?

The access to full-length viral genomes has highlighted the major role of recombination in the evolution of RNA and DNA plant viruses (reviewed in 84, 138).

Recombination is known to blur phylogenetic signals; thus, ignoring it when reconstructing the evolutionary histories of viruses will likely lead to misleading inferences (127). It is therefore highly desirable to either exclude recombinant sequences or focus the analysis on nonrecombined genomic regions. Numerous methods and computer programs have been developed for detecting recombination and locating recombination breakpoints (see **Supplemental Table 1**) (85). Their relative performance in terms of power (probability to detect true recombination events) and specificity (avoidance of false positives) has been assessed (107).

However, accounting for recombination may provide extra information to infer transmission trees. Indeed, such evolutionary events occur during multiple infections, which imply that viruses with potentially different geographical origins have simultaneously shared the same host. Nonrandom patterns of sequence exchanges may also provide valuable information about potential geographical or ecological barriers (72). Despite methodological developments such as ancestral recombination graphs (98), computational and theoretical obstacles remain before we can truly integrate recombination in phylodynamic inference (43).

and/or particular recombination breakpoints (12, 120). Moreover, whole-genome sequencing offered unprecedented insights into the infraspecific genetic polymorphism and further evidenced the major role of recombination and reassortment in plant virus evolution (100, 138) (see sidebar titled Recombinant Genomes: Trouble or Treasure?). However, Sanger sequencing approaches have limited throughput, are both resource- and labor-intensive, and depend upon a priori knowledge of virus sequences. Moreover, polymerase-based techniques may be error-prone depending on the enzyme used. Thus, the corresponding data may not always reflect the true viral genetic diversity (94).

With the recent advent of high-throughput sequencing (HTS) technologies, the ability to generate large amounts of sequence data at relatively low cost led to breakthroughs in plant virus discovery and molecular epidemiology. Because they require little a priori knowledge of the targeted virus, metagenomic approaches have enabled the identification of hundreds of unknown viruses (17, 104, 123) as well as the discovery of new variants of known virus species that escaped existing detection procedures (83). Such approaches will undoubtedly improve our understanding of the distribution and dynamics of plant virus diversity in both cultivated and natural areas (123). In addition, HTS technologies can be used to generate consensus genome sequences without an amplification step (79) or a deep characterization of within-host diversity (20, 133).

Reassortment: exchange of full genomic segments resulting in infectious units with new combinations of segments

HTS: high-throughput sequencing

Figure 1

Evolutionary processes leading to the viral diversity observed in a heterogeneous landscape. Different steps of the infection process are presented, from the inoculation of a single cell by two different viral genomes (*orange and blue*) to the systemic infection of the host. Virus replication is an error-prone process that results in the diversification of the viral genetic material through mutation (*white and black bars*) and recombination events (orange/blue chimeric genomes). A population of viral genomes is therefore generated during infection. However, the selection of fitter individuals at the cellular level combined with bottlenecks occurring during host colonization reduces the range of genetic variation within the infected host. Additional bottlenecks during plant-to-plant transmission (generally via vectors) lead to the efficient inoculation of a limited number of virus genomes. The epidemiological processes then shape the viral population according to host features. Viral populations can be characterized using serological [e.g., enzyme-linked immunosorbent

New sequencing technologies, including single-molecule real-time (SMRT) sequencing and other long-read sequencing technologies, should also provide solutions for real-time genomic surveillance of viral outbreaks in the next few years (16, 112).

plotype: unique nbination of rkers on a haploid nome

Sampling Design to Measure Virus Prevalence and Diversity

Most molecular epidemiology approaches require assessing and comparing the genetic diversity of viral populations sampled from different hosts at different spatial (possibly from the host up to the continent) and temporal (often multiple years up to several decades) scales. The diversity of viral populations can be estimated using different criteria according to the research questions, the type of genetic data obtained (i.e., molecular markers targeting one or several genomic regions, partial- or whole-genome sequences), and the analytical method chosen (44). Classical approaches aim at assessing the number and frequency of different haplotypes and the genetic distances between and within populations (92). Besides providing a direct estimation of genetic distances, partial- and whole-genome sequences also enable the quantification of the effects of different evolutionary forces, demographic inference, and the reconstruction of genealogical or phylogenetic relationships.

Biological and environmental variables (e.g., the life cycle of hosts and vectors, host/nonhost crop rotations, landscape structure, dispersal distances of vectors, etc.) that can impact epidemics should be considered when designing appropriate sampling schemes. The presence of symptoms can be used to target infected hosts, but when the study implies assessing relative virus prevalence (e.g., of different strains or in different hosts), plants should be collected regardless of symptom expression to avoid bias due to tolerance or asymptomatic stages of infection. Asymptomatic infections are not predominant in the cultivated compartment (except for tolerant host varieties), but they can represent an important proportion of plants in the wild compartment (122). Moreover, the type of plant material (e.g., leaves or stalks) collected during surveys has to be carefully considered, particularly for samples from the wild compartment, because virus concentration can be low and heterogeneous in infected plants (73). As most plant viral species are transmitted by vectors, sampling of plant material can be advantageously completed by collecting insects from which the virus can be extracted and sequenced (97). Indeed, the comparison between viral lineages found in insects and plants can provide information on the epidemiological cycle and dispersal of the virus. Sampling design and effort should also be adapted to the aim of the study. If intensive sampling of infected hosts is generally required to reconstruct transmission chains (see section How Did the Pathogen Spread within an Outbreak?), less intensive but well-balanced sampling (111) can be sufficient to describe viral diversity, compare population structures, and reconstruct dispersal and introduction events. Rarefaction curves and nonparametric richness estimators can be used to adapt sampling efforts and compare genetic diversities (53, 61). Moreover, hierarchical sampling and hierarchical partitioning of samples among variation factors may allow testing their effect on plant virus genetic differentiation (34). The following sections present the main approaches used to analyze viral sequence data to uncover the spatiotemporal dynamics of plant virus epidemics from the continental scale to the single outbreak.

HOW DO VIRUSES INVADE NEW TERRITORIES AND FURTHER SPREAD IN LANDSCAPES?

During the past few decades, trade globalization and greater human mobility have largely con-

are key to preventing new introductions and improving management strategies (47). Given the complex nature of spatiotemporal interactions across multiple scales, determining and managing the key processes driving pathogen dispersal are challenging. Although an appropriate scale for data collection and analysis should match the scale of the ecological phenomenon under question (87), multiscale information may be necessary to gain a more holistic view of transmission dynamics.

Genetics-based methods to study the spread of pathogens typically stem from the complementary fields of population genetics, landscape genetics, and phylogeography (8, 118, 144). These disciplines generally differ not only in terms of data and analyses commonly used but also by the timescale over which the data are informative. Indeed, population genetics and more recent landscape genetics approaches often use neutral genetic markers to infer population structure and contemporary gene flow at local or regional spatial scales (52). In contrast, phylogeography is mainly based on sequence data and aims to reconstruct long-term population dynamics (usually at an evolutionary timescale) such as dispersal events at continental or global scales (3). However, for measurably evolving pathogens, phylogeographic methods can also reveal patterns at spatial and temporal scales usually investigated using population and landscape genetics approaches (8).

Today, major advances in genetic and spatial data acquisition tools alongside subsequent analytical methods provide new opportunities to infer, within formal statistical frameworks, the processes at the origin of the spatial distributions of viruses and to quantitatively evaluate potential predictors of spread in complex environmental settings (8, 111) (see **Supplemental Table 1**).

Population genetics: study of genetic variation within and between populations

Landscape genetics: study of the geographical and environmental features that structure genetic variation (combines landscape ecology and population genetics)

Phylogeography: study of the spatiotemporal distribution of genetic lineages

Supplemental Material

Exploratory Approaches

Many of the population genetics methods developed to describe spatial genetic structures and estimate migration parameters require neutral and independent markers and/or rely on equilibrium assumptions (Hardy-Weinberg equilibrium, linkage equilibrium) that are rarely met for viruses (49, 51). Up to a decade ago, epidemiological studies focusing on describing and comparing the genetic structure of plant virus populations used molecular markers and/or partial genomic sequences to compute various indices of genetic diversity and measures of differentiation, e.g., mean pairwise nucleotide differences, number of polymorphic sites, pairwise genetic distances, and statistics of differentiation such as F_{ST} and K_{ST} (reviewed in 92). Further developments of versatile software such as Arlequin (37), which implements hierarchical analysis of molecular variance (AMOVA), have provided useful approaches to test for population subdivision according, for example, to geography or host plant species (34, 99, 108). The AMOVA design requires hypotheses on the genetic structure to be tested (e.g., samples are grouped according to geography or host plants). Clustering analyses that do not require such hypotheses on the structuring factors (i.e., a priori characterization of genetic groups) can thus be more appealing to analyze subdivisions in virus populations and identify immigrant genotypes (100, 108, 147). Most model-based clustering methods aim to maximize Hardy-Weinberg and linkage equilibria. Thus, when used on virus data, which are likely to deviate from these assumptions to various degrees, results should be carefully interpreted and completed with some kind of robustness analysis (145). Alternatively, although rarely used on plant viruses, exploratory methods that do not rely on genetic models constitute a valuable first step to assess both spatial and temporal structures of the genetic diversity within virus populations, as well as genotype flow between host populations (65). For example, spatial analysis of molecular variance (SAMOVA; 33) combined with Monmonier's maximum-difference algorithm (90) allowed the identification of both genetic subgroups and major disruptions of geno-

pairwise information (MAPI)], which provides spatial maps of mean genetic differentiation estimated between virus sequences (106). Multivariate analyses, such as the discriminant analysis of principal components (DAPC), can also be used on virus genetic data to analyze population substructure, perform probabilistic assignments (i.e., to detect immigrating genotypes), and identify the most important mutations involved in differentiation between genetic groups (65). De Bruyn et al. (22) combined spatial principal components analyses (sPCAs; 66) and DAPC (65) to study the spatial genetic structure of geminiviruses causing cassava mosaic disease in Madagascar. Such flexible exploratory methods are especially interesting, as they ease the processing of the increasingly large data sets generated using HTS technologies, and they are relatively easy to apply using packages [e.g., adegenet (66) and poppr (69)] of the R statistical software (113).

yesian inference: tistical inference thod in which ves' rule is used to ovide probability tributions of model rameters

RCA: time to the st recent common cestor

terochronous quence: dated netic sequences mpled at different nts in time

Reconstructing Invasion Pathways

Several phylogeographic frameworks are available to infer ancestral locations and spatiotemporal dynamics. These approaches mainly differ in their ability to handle spatial information (separately or simultaneously with the phylogenetic reconstruction) and to account for uncertainty (9). Recent approaches targeting viruses are based on the reconstruction of phylogenies in which temporal and spatial information are explicitly integrated to allow for the simultaneous inference of these processes (28, 77, 78). Moreover, statistical parametric or nonparametric models based on coalescent theory can be used to directly link patterns of genetic diversity to the demographic history of viral populations in a phylodynamic framework (50, 109). The popular programs BEAST and BEAST2 offer an integrative platform to perform these analyses (10, 30). Using a Bayesian inference framework for testing evolutionary hypotheses while accounting for phylogenetic uncertainty, they integrate numerous molecular clock models, discrete and continuous diffusion, and population dynamics. Although being largely validated and used on data sets of human and animal viruses, these methods have been applied only recently to RNA (21, 101, 114, 140, 148), ssDNA (1, 22, 23, 75, 80, 89, 137), and double-stranded (dsDNA) (147) plant viruses. Given their high potential, phylogeographic analyses are likely to keep gaining popularity in plant virus molecular epidemiology studies in the coming years. Here, we describe more precisely the data and methods required to address the questions relative to the geographical origin of a given viral lineage and the reconstruction of invasion pathways.

Phylogeographic analyses commonly use molecular clock models to represent the relationship between genetic distance and calendar time. Consequently, this can be used to estimate the ages of branching events, including the time to the most recent common ancestor (tMRCA) of lineages of interest. Many molecular clock models are available to accommodate for rate heterogeneities (58). Although initial strict clock models assumed a constant rate of molecular evolution throughout the tree, relaxed clocks now allow branch-specific evolutionary rates (27, 58). To calibrate such molecular clocks, studies targeting measurably evolving pathogens such as viruses use heterochronous sequence data. An evolutionary rate can thus be estimated, usually given as a number of nucleotide substitutions per site per year (29). Appropriate temporal sampling allowing the accumulation of genetic variation is recommended to enhance the temporal signal, whereas long-enough genomic sequences are necessary to increase the phylogenetic resolution (111, 130). The use of a herbarium or archeological specimens may allow for a greater temporal depth and thus more precise evolutionary estimates (82, 134). The presence of a temporal signal in the data set should always be tested; different methods and programs can be used (117), including linear regression of phylogenetic root-to-tip distance against sampling date and date-randomization tests (31). However,

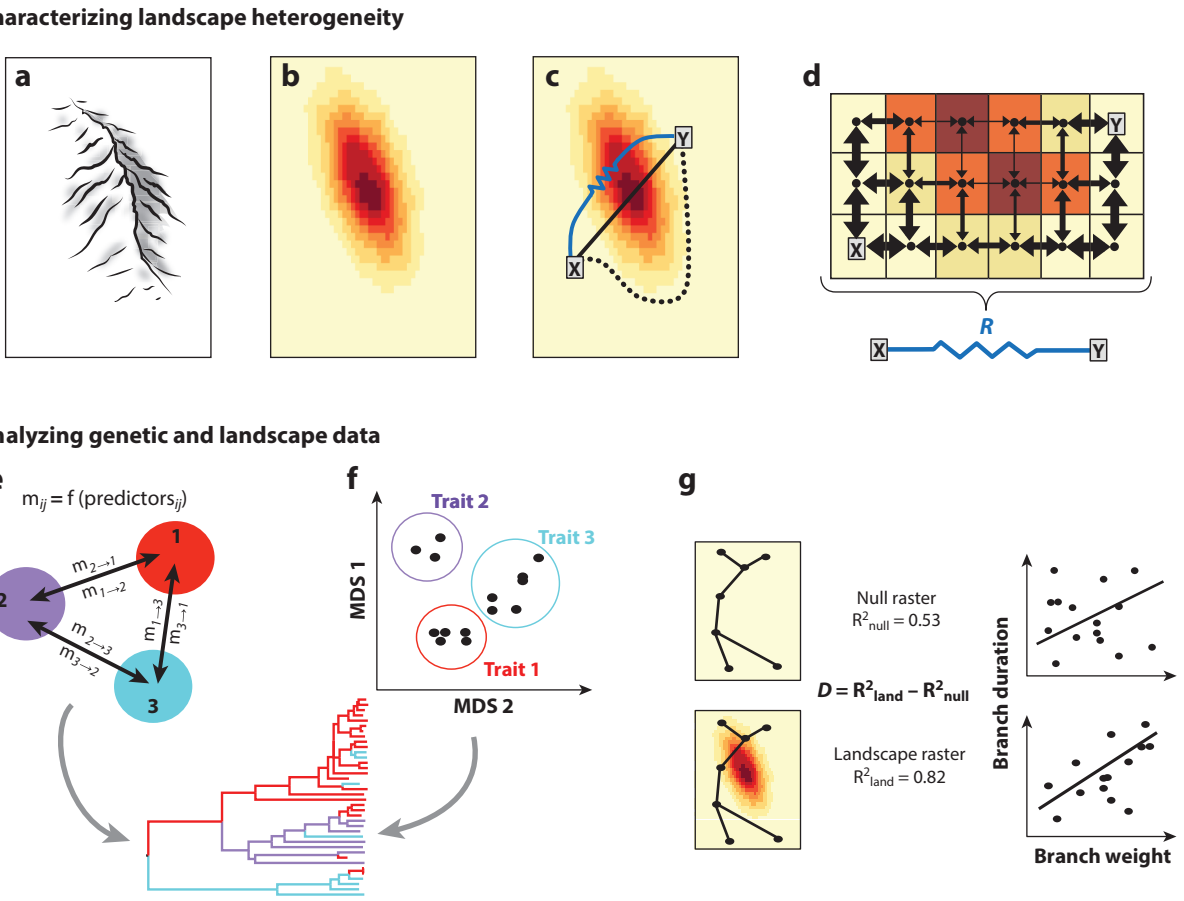
Besides inferring emergence or introduction dates of a given viral lineage in a new location (22, 114, 140), time-calibrated phylogenies can be used to evaluate the efficiency and timeliness of an epidemiological surveillance system by comparing the estimated MRCA ages with the dates of discovery of a given outbreak (114).

Using Bayesian skyline plots and other coalescent-based methods, it is also possible to estimate effective population sizes through time (59) and detect population bottlenecks and subsequent expansion during invasions (1). However, these methods often assume a single well-mixed population, an assumption that is only rarely met as virus populations may be highly structured (spatially and/or by host). Because violation of this assumption can lead to misleading inference (57), a cautious interpretation is required.

Popular phylogeographic methods for reconstructing virus spatial spread from genetic data treat the geographical locations assigned to each sequence as discrete traits (e.g., for viruses sampled in cities or countries) or continuous traits (e.g., samples with latitude-longitude coordinates). Movements are represented as changes in traits along sampled lineages (77, 78). In discrete phylogeography, stochastic diffusion processes are modeled using a continuous-time Markov chain (CTMC), where the transitions between spatial locations in the phylogeny are either symmetrical or asymmetrical to provide a more realistic description of the spatial dynamics. The number of transitions between spatial locations can be inferred, providing valuable information when one is interested in the number and direction of migration events in source-sink dynamics (77). The most significant dispersal pathways can then be identified using Bayesian stochastic search variable selection (BSSVS) (77). The continuous diffusion model relies on relaxed random walk models (Brownian motion process) to explore two-dimensional space and can yield more realistic reconstructions of the dispersal process in a given landscape (78). These models enable the computation of various statistics to quantify the spatial dynamics of an epidemic, such as the diffusion coefficient D that measures spatial velocity (50). Both approaches have been used to reconstruct the invasion pathways of various plant viruses at global and regional scales (75, 137, 140). It is important to emphasize that the accuracy of these methods in estimating the location of ancestors and capturing dispersal patterns is directly linked to the quality of sampling (142). Estimation of ancestral locations might be highly uncertain if an inferred ancestor is only distantly related (spatially) to the sampled cases. Moreover, if samples from key locations or regions are absent or rare, then virus movements will be underestimated and the inferred locations of ancestors may be biased toward over-represented locations. Although these methods are particularly efficient from a computational perspective, a recent study has provided evidence that they may suffer from various biases and statistical inefficiency (81). A new model-based approach, Bayesian structured coalescent approximation (BASTA), has been developed (81) and is implemented in BEAST2. This method is based on the structured coalescent, a statistical model that explicitly accounts for migration effects on the shape and branch lengths of the genealogy.

Phylogeographic approaches have benefited from a rich development of statistical inference tools. The successful application of these methods to plant viruses depends on the assembly of large collections of dated and georeferenced plant virus sequences, as are already available for numerous human/animal viruses. However, despite the huge number of plant virus sequences in molecular databases, temporal and spatial information associated with the submitted sequences are often lacking for the sequences older than the past decade (22).

Integrating Landscape Heterogeneity



gure 2

lating genetic patterns to landscape characteristics. (*Top*) Characterizing landscape heterogeneity using resistance surfaces: (*a*) an mple in which mountains are dispersal barriers; (*b*) rasterized resistance surface in which darker cells indicate higher resistance to persal; (*c*) possible distance metrics between focal points X and Y: Euclidean distance (*solid black line*), least-cost path (*dotted line*), and esistance distance *R* across multiple pathways (*blue line*); (*d*) details of the calculation of *R* between X and Y with heavier arrows icating increasingly facilitated virus flow. (*Bottom*) Methods to incorporate landscape heterogeneity into (*e*, *f*) discrete and continuous phylogeographic analyses: (*e*) generalized linear model (GLM) extension of the diffusion model in which diffusion rates $_{ij}$) between demes (*colored circles*) are a function of a set of explanatory variables tested using Bayesian model averaging (76); pathogen spread relative to a null model, quantified after assignment of phylogeographic traits through landscape-informed stering (here, three clusters) of pathogen locations via multidimensional scaling (14); (*g*) assessment of the effect *D* of an vironmental variable through an increase in the association between lineage movements (branch duration, inferred from a dated ylogeny) and the associated resistance weights, relative to a null model (24).

host distributions, vector movements, and the transfer of infected plant material (87). With the modern capacity to produce and analyze genetic data, new opportunities have arisen to use molecular epidemiology analyses to gain a detailed quantitative understanding of these interactions (**Figure 2**).

ntel test: tistical test of the rrelation between

The effect of landscape heterogeneity on transmission is difficult to quantify given the un-

structure and landscape variables, particularly for barrier effects (65, 145). However, the suitability of such tests to detect landscape effects has come into question (5). Simulation models offer an alternative quantitative approach, an example being the identification of a 50% permeability of rivers to raccoon movements and thus to rabies virus in North America (116). Furthermore, with the accessibility of finely resolved genetic data, the analyses now exploit sophisticated Bayesian phylogeographic frameworks (see section Reconstructing Invasion Pathways) to measure variation in dispersal among landscape components (77, 78). However, it is only recently that a statistical framework, utilizing a generalized linear model (GLM) parameterization, has become available to simultaneously test and quantify the effects of potential predictors on dispersal patterns (41, 76). This has identified the role of human and animal transportation networks on influenza spread (76, 96) and has recently been applied for the first time to a plant virus (140).

An increasingly popular method to account for landscape heterogeneity is to represent variables in terms of their cost or "resistance" to dispersal, based on the "isolation by resistance" (IBR) concept (86). Computer programs such as Circuitscape (131) and the R package gdistance (141) provide various distance-based metrics to measure dispersal potential across different landscape resistance surfaces. Synthesizing landscape information in this way provides a simple input for modeling approaches, as exemplified by the use of a resistance surface based on rice production statistics to quantify the impact of crop intensification on *Rice yellow mottle virus* (RYMV) spread (140). Several other IBR approaches have emerged, including a method to compare phylogenetic reconstructions of dispersal with landscape variation (24). This work provides a framework to extract information from the branches of spatiotemporally referenced phylogenies to perform tests of correlations with landscape characteristics, employing a randomization procedure to determine significance. The framework offers some flexibility in terms of the method and software used to build phylogenies and is less computationally demanding than the GLM approach (76). However, reliance on linear regression to identify correlations may not capture more complex relationships (e.g., quadratic and thresholds) between dispersal and landscape features. Alternatively, Brunker et al. (15) used resistance distances to rescale spatial information and assign phylogenetic traits as a means to directly inform phylogeographic reconstructions, simultaneously providing a means to test the effect of landscape features on epidemics across multiple spatial scales.

These promising ways to integrate landscape heterogeneity are still under development. Potential future improvements include the use of nonlinear multivariate approaches, development of simulation models to assess the relative sensitivity of the various methods to detect barrier effects, and exploration and integration of the temporal dynamics of landscape heterogeneity. More generally, the application of phylodynamic techniques to identify important sources of variation in dispersal is a potentially fruitful endeavor for the next few years (4).

HOW DID THE PATHOGEN SPREAD WITHIN AN OUTBREAK?

Another field which has been developing at an ever-increasing pace during the past decade is the reconstruction of the transmission links within outbreaks. Inferring the history of transmission events within a host population can highlight key drivers of transmission, provide refined estimates of epidemiological parameters and point out risk factors related to vectors, reservoirs, and landscape components (103). Ultimately, such studies can help build epidemiological projections, design control strategies, and deliver scientific advice to governmental agencies. However, inferring "who infected whom" in outbreaks of infectious diseases remains a challenging task.

IR: susceptible, posed, infectious, moved

outbreaks (19, 56, 64, 67, 88, 91). Data can be epidemiological records, such as the spatiotemporal locations of infected hosts, or genetic information on evolutionary relationships between virus genomes sampled from the hosts. In particular, when enough mutations can be observed during an outbreak, the joint analysis of epidemiological and genetic data can provide valuable insights into transmission dynamics. Several approaches currently under development aim to appropriately combine these data. Specifically, we highlight the reconstruction of transmission trees and the estimation of epidemiological parameters. We present below the existing approaches that address this question to determine how measurably evolving pathogens spread within a host population.

Model-Based Inference of Transmission Trees

Transmission trees have been inferred using various modeling approaches. Some of these approaches are intrinsically based on phylogenetic models in which epidemiological information is introduced. Others have started with epidemiological models enriched with genetic information.

The first approach is based on phylogenetic and coalescent models (117). Here, spatial or temporal information is added to the process of phylogenetic reconstruction. Such methods relate the demography of the pathogen to its evolution and may incorporate a diffusion model to account for the movement of the pathogen over geographical space (50, 54, 77, 78, 110, 115, 132) (see section Reconstructing Invasion Pathways). This approach is relatively robust to the intensity of epidemiological sampling. However, because the underlying models do not have an explicit epidemiological formulation (except for some models; 54), the inferred parameters cannot be easily related to the epidemiological processes. Jombart et al. (67) also pointed out that a phylogenetic approach attempts to infer hypothetical common ancestors among the sampled genomes and thus may not be appropriate for a set of genomes containing both ancestors and their descendants. Indeed, phylogenetic methods consider that sampled strains are all tips of an unknown genealogy, making it impossible for a sampled strain to be (directly or indirectly) the ancestor of another sampled strain (121), an issue that is often encountered for densely sampled outbreaks. However, recent works have addressed this issue by developing an algorithm to infer phylogenetic trees in which sampled sequences can be direct ancestors of other sampled sequences (45) or by employing an individual-based disease transmission model and a coalescent process taking place within each host (54).

The second approach uses spatial epidemiological models of transmission and models of genetic drift to directly reconstruct the transmission tree reflecting "who infected whom." This approach is generally based on stochastic and spatiotemporal SEIR (susceptible, exposed, infectious, removed) models explicitly representing successive states of host individuals (64, 88, 91, 135, 149, 150) to recognize the host population structure and epidemiological processes governing host-pathogen interactions. A model for the spread of the pathogen in the population and a model for the accumulation of point mutations over time are often used to calculate the probability that the genetic sequence transmitted from case A to case B could have mutated into the sequences sampled from the two cases in the duration between transmission and sampling (**Figure 3**) (88). More specifically, a first study identified a large set of transmission trees consistent with the available genetic data and then ranked these trees with respect to a likelihood computed from temporal data, revealing the most likely set of transmission trees (19). In later works, the likelihood of the transmission tree J given temporal (T), spatial (X), and genetic (G) data was approximated by the product of three independent likelihoods (150): $L(J|T,X,G) = L1(J|T) \times L2(J|X) \times L3(J|G)$.

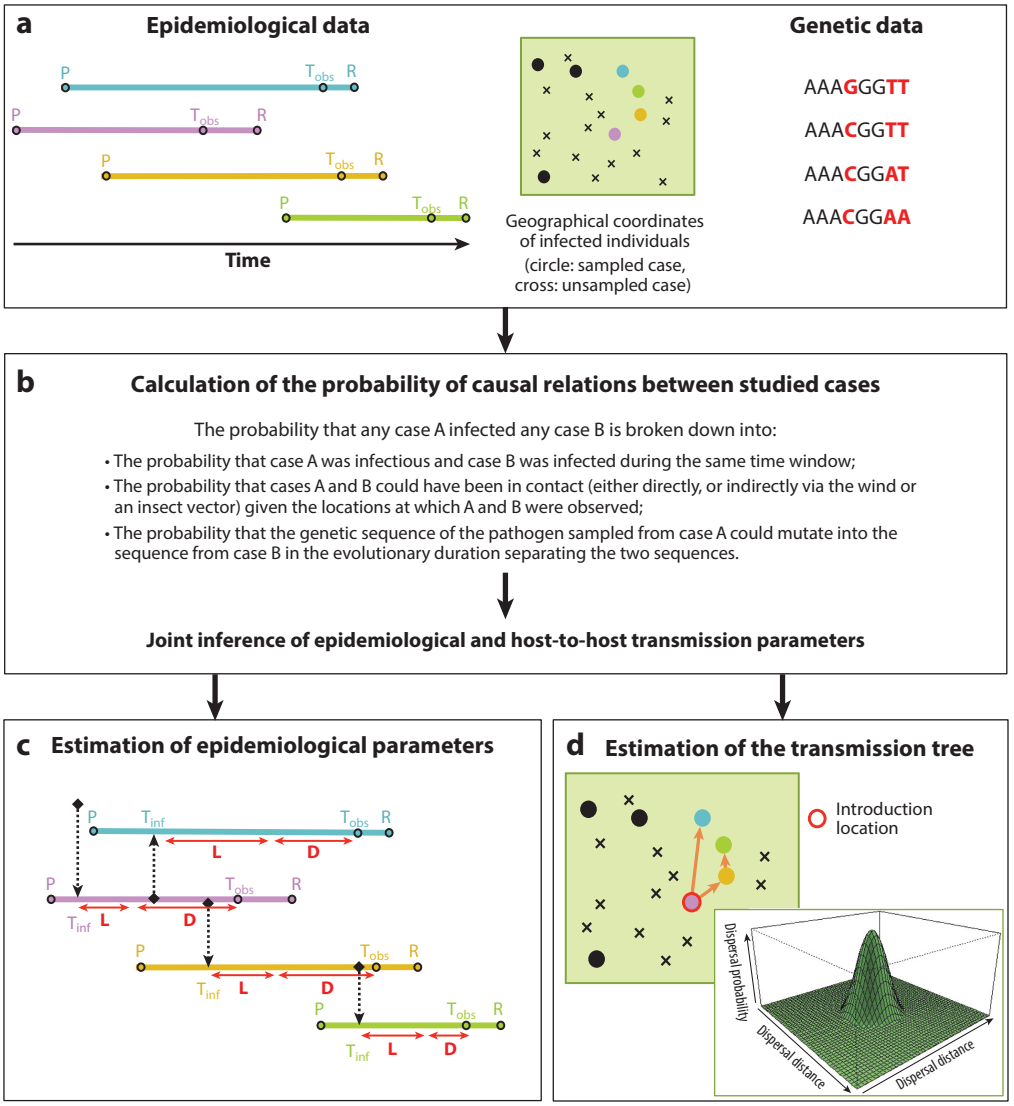


Figure 3

Inference of epidemiological parameters and transmission trees in a landscape. (*a*) Inputs of space-time-genetic SEIR (susceptible, exposed, infectious, removed) models are epidemiological data (e.g., plantation dates, dates of symptom detection, removal dates, locations of infected trees) and genetic data (genetic sequences). (*b*) Based on these inputs and the calculation of probabilities of causal relations between studied cases, the space-time-genetic SEIR models and the accompanying estimation algorithms provide joint estimations of transmission links and epidemiological parameters. (*c*) Estimation of epidemiological parameters defining the duration of the latent period, the duration of the infectious period before detection, and the time of infection. (*d*) Estimation of the transmission tree ("who infected whom") thanks to spatial parameters (introduction location and dispersal function). Abbreviations: P, planting date; T_{obs}, time of symptom detection; R, removal date; L, duration of the latency period; D, duration of the infectious period before detection; T_{inf}, time of infection.

into account the inherent dependence between temporal, spatial, and genetic data and calculated the likelihood of transmission trees (88, 91). These methods have been very valuable in unraveling transmission pathways during outbreaks. However, they either avoid explicit inference of the unobserved pathogen sequences transmitted during infection (64, 88, 91, 149, 150) or use approximate Bayesian inference to account for these sequences (135). Such approximate approaches greatly reduce the computational challenges associated with inferring the unobserved transmitted sequences and facilitate statistical inference, particularly when the transmission tree is of primary interest. Instead of using approximations, Lau et al. (74) considered a Bayesian framework that simultaneously and explicitly infers the transmission tree and the transmitted pathogen sequences. This approach facilitates the use of realistic likelihood functions and allows the systematic joint inference of epidemio-evolutionary processes from partially observed outbreaks.

Compared with phylogenetic approaches, space-time-genetic SEIR approaches generally require a moderate to high proportion of infected hosts for accurate inference. This is particularly true for early studies assuming that sampled cases were directly related through transmission (19, 67, 91). More recent works accommodate the inherent complexities of polyphyletic and partially sampled outbreaks (64, 74, 88). Thus, space-time-genetic SEIR models and associated estimation algorithms yield increasingly satisfactory reconstructions of transmission trees. Such models can nevertheless result in misleading interpretations of transmission dynamics if they use a single sequence from each infected case in situations of mixed infections (25, 146), although works in progress tend to overcome this problem (25). Phylogenetic and space-time-genetic SEIR approaches have recently begun to merge by combining features of phylogenetic and transmission tree approaches to reconstruct partially observed transmission networks (35, 39, 54, 70). In addition to their ability to infer a transmission tree, most of the approaches presented in this section provide estimates of other important epidemiological parameters.

sic reproduction mber: number of ividuals infected ring the infectious iod of an infected ividual placed into uninfected pulation

tent period: delay ween inoculation of athogen and the ginning of the host ectious period

cubation period: ay between culation of a hogen and the set of the first nptom of the ected host

spersal function: obability tribution of the ival location of an ividual (e.g., an ect vector) persed from a given rting location

Estimation of Epidemiological Parameters

For several decades now, S(E)I(R) models (without genetics) have been fitted to data on the number of cases through time to estimate epidemiological parameters such as the basic reproduction number R_0 or thresholds of vaccination coverage (2, 13, 26). The emergence of epidemio-evolutionary approaches based on S(E)I(R) models should lead to finer estimations by exploiting information brought by genetic data. These approaches often explicitly include (and allow inference about) parameters related to infection strength (and sometimes its heterogeneity among hosts), the latent period, the incubation period, the dispersal function (which partly determines the speed and spatial extent of disease spread), and the substitution rate (e.g., 136). In addition, these approaches allow the calculation of the effective reproduction number over time or the total infected population over a given spatiotemporal window.

The estimation algorithms have generally been developed within a Bayesian framework (25, 54, 88, 91, 135, 150) to incorporate prior knowledge about the parameters and to benefit from techniques allowing the inference of hidden variables, such as infection times and transmitted pathogen sequences when transmission trees and epidemiological parameters are estimated jointly. It is especially interesting to incorporate prior knowledge about influential epidemiological or evolutionary parameters on which the data used for model fitting bring little information. Finally, the output of Bayesian estimation algorithms is a sample of the joint posterior distribution of the parameters. Such samples can be used to provide not only point estimates of parameters but also uncertainties in, and dependencies between, estimates. Such models have not yet been used to

SUMMARY POINTS

1. The high substitution rate of viruses implies that evolutionary and epidemiological processes are observable at the same timescales, and that viral genomes are scattered with imprints that can be used to infer virus dynamics in landscapes through space and time.
2. Continual advances in virus characterization methods have vastly expanded our knowledge of the existing virus species and of their intraspecific diversity.
3. Appropriate sampling schemes are required to prevent bias when studying how the diversity of viral populations is structured by biological and environmental variables.
4. Because viruses are clonal, assumption-free exploratory analyses are more appropriate than classical population genetics approaches to describe the spatial structure of viral diversity.
5. Phylogeographic models enable inference of invasion pathways over large areas based on the geographical coordinates of dated sequences.
6. New approaches combining landscape genetics and phylogeography provide a means to test the impact of landscape configuration and composition on virus spatiotemporal dynamics.
7. Recent phylodynamic and space-time-genetic SEIR models can be used to infer transmission trees and other key epidemiological and evolutionary parameters, based on virus sequences from intensively sampled outbreaks.

FUTURE ISSUES

1. Characterizing plant virus diversity at the ecosystem scale is still needed to better understand the spatiotemporal dynamics of plant viruses in cultivated and natural areas.
2. Molecular epidemiology studies should considerably benefit from advances in real-time, portable genome sequencing and high-throughput sequencing to produce long reads and high-fidelity sequences.
3. In parallel, more powerful estimation approaches will be welcome to exploit the ever-increasing number of sequences representing virus diversity both between and within hosts. Progress could take the form of faster algorithms using robust approximations, more flexible models, and complex models of the various processes underlying within- and between-host dynamics.
4. Molecular epidemiology studies generally focus on the neutral genetic diversity of non-recombinant sequences. Integrating information brought by recombinant sequences and relating genetic changes under selection with epidemiological changes are promising methodological challenges.
5. Better characterizing the various landscape types and host characteristics and estimating their impact are both challenging and important for the understanding of plant virus spread.
6. High-resolution inference of "who infected whom" based on sequencing data is a promis-

7. Development of new frameworks to enable improved integration of data and models may lead to real-time characterization and prediction of outbreaks. This might take the form of streamlined pipelines from sample collection to sequencing, from bioinformatics analysis through updated phylogenies to estimation of parameters feeding disease management models and, finally, feedback procedures toward disease control organizations.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work has been supported by an INRA/ANSES PhD grant (C.P.), the EU FP7-PEOPLE program (PIOF-GA-2013-622571) (P.R.), the ANR-funded SMITID project (ANR-16-CE35-0006), and Agropolis Fondation (E-SPACE project).

LITERATURE CITED

1. Almeida RP, Bennett GM, Anhalt MD, Tsai CW, O'Grady P. 2009. Spread of an introduced vector-borne banana virus in Hawaii. *Mol. Ecol.* 18(1):136–46
2. Anderson RM, May RM. 1992. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford Univ. Press
3. Avise JC. 2000. *Phylogeography: The History and Formation of Species*. Cambridge, MA: Harvard Univ. Press
4. Baele G, Suchard MA, Rambaut A, Lemey P. 2017. Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.* 66(1):e47–65
5. Balkenhol N, Waits LP, Dezzani RJ. 2009. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* 32(5):818–30
6. Bebber DP, Holmes T, Gurr SJ. 2014. The global spread of crop pests and pathogens. *Glob. Ecol. Biogeogr.* 23(12):1398–407
7. **Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* 30(6):306–13**
8. Biek R, Real LA. 2010. The landscape genetics of infectious disease emergence and spread. *Mol. Ecol.* 19(17):3515–31
9. Bloomquist EW, Lemey P, Suchard MA. 2010. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.* 25(11):626–32
10. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, et al. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* 10(4):e1003537
11. Bousalem M, Dallot S, Fuji S, Natsuaki KT. 2003. Origin, world-wide dispersion, bio-geographical diversification, radiation and recombination: an evolutionary history of *Yam mild mosaic virus* (YMMV). *Infect. Genet. Evol.* 3(3):189–206
12. Bousalem M, Dallot S, Guyader S. 2000. The use of phylogenetic data to develop molecular tools for the detection and genotyping of *Yam mosaic virus*. Potential application in molecular epidemiology. *J. Virol. Methods* 90(1):25–36
13. Britton T, Giardina F. 2016. Introduction to statistical inference for infectious diseases. *J. Société Fr.*

This review highlights landscape genetic principles relevant to infection dynamics.

15. Brunker K, Marston DA, Horton DL, Cleaveland S, Fooks AR, et al. 2015. Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing. *Virus Evol*. 1(1):vev011
16. Bull RA, Eltahla AA, Rodrigo C, Koekkoek SM, Walker M, et al. 2016. A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genom*. 17:247
17. Candresse T, Filloux D, Muhire B, Julian C, Galzi S, et al. 2014. Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLOS ONE* 9(7):e102945
18. Clark MF, Adams AN. 1977. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *J. Gen. Virol.* 34(3):475–83
19. **Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, et al. 2008. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* 275(1637):887–95**
20. Cuevas JM, Willemsen A, Hillung J, Zwart MP, Elena SF. 2015. Temporal dynamics of intrahost molecular evolution for a plant RNA virus. *Mol. Biol. Evol.* 32(5):1132–47
21. Davino S, Willemsen A, Panno S, Davino M, Catara A, et al. 2013. Emergence and phylodynamics of *Citrus tristeza virus* in Sicily, Italy. *PLOS ONE* 8(6):e66700
22. De Bruyn A, Harimalala M, Zinga I, Mabvakure BM, Hoareau M, et al. 2016. Divergent evolutionary and epidemiological dynamics of cassava mosaic geminiviruses in Madagascar. *BMC Evol. Biol.* 16:182
23. De Bruyn A, Villemot J, Lefeuvre P, Villar E, Hoareau M, et al. 2012. East African cassava mosaic-like viruses from Africa to Indian Ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. *BMC Evol. Biol.* 12:228
24. Dellicour S, Rose R, Pybus OG. 2016. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinform*. 17:82
25. Didelot X, Gardy J, Colijn C. 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* 31(7):1869–79
26. Diekmann O, Heesterbeek H, Britton T. 2012. *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton, NJ: Princeton Univ. Press
27. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLOS Biol*. 4(5):e88
28. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161(3):1307–20
29. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends Ecol. Evol.* 18(9):481–88
30. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29(8):1969–73
31. Duchêne S, Duchêne D, Holmes EC, Ho SYW. 2015. The performance of the date-randomisation test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* 32(7):1895–906
32. **Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9(4):267–76**
33. Dupanloup I, Schneider S, Excoffier L. 2002. A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.* 11(12):2571–81
34. D'Urso F, Sambade A, Moya A, Guerri J, Moreno P. 2003. Variation of haplotype distributions of two genomic regions of *Citrus tristeza virus* populations from eastern Spain. *Mol. Ecol.* 12(2):517–26
35. Eldholm V, Rieux A, Monteserin J, Lopez JM, Palmero D, et al. 2016. Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *eLife* 5:e16644
36. Elena SF, Bedhomme S, Carrasco P, Cuevas JM, de la Iglesia F, et al. 2011. The evolutionary genetics of emerging plant RNA viruses. *Mol. Plant-Microbe Interact.* 24(3):287–93
37. Excoffier L, Laval G, Schneider S. 2007. An integrated software package for population genetics data analysis. *Evol. Bioinform.* 1:47–50
38. Fabre F, Moury B, Johansen EI, Simon V, Jacquemond M, Senoussi R. 2014. Narrow bottlenecks affect

19. This early approach infers transmission trees using epidemiological data and virus genomes.

32. This review clearly explains the factors affecting virus mutation and substitution rates.

39. Famulare M, Hu H. 2015. Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. *Int. Health* 7(2):130–38
40. Fargette D, Konaté G, Fauquet C, Muller E, Peterschmitt M, Thresh JM. 2006. Molecular ecology and emergence of tropical plant viruses. *Annu. Rev. Phytopathol.* 44:235–60
41. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. 2013. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos. Trans. R. Soc. B* 368(1614):20120196
42. Fraile A, Malpica JM, Aranda MA, Rodríguez-Cerezo E, García-Arenal F. 1996. Genetic diversity in tobacco mild green mosaic tobamovirus infecting the wild plant *Nicotiana glauca*. *Virology* 223(1):148–55
43. Frost SDW, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. 2015. Eight challenges in phylodynamic inference. *Epidemics* 10:88–92
44. García-Arenal F, Fraile A, Malpica JM. 2001. Variability and genetic structure of plant virus populations. *Annu. Rev. Phytopathol.* 39:157–86
45. Gavryushkina A, Welch D, Stadler T, Drummond AJ. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLOS Comput. Biol.* 10(12):e1003919
46. Gibbs AJ, Fargette D, García-Arenal F, Gibbs MJ. 2010. Time—the emerging dimension of plant virus studies. *J. Gen. Virol.* 91(1):13–22
47. Gilligan CA. 2008. Sustainable agriculture and plant diseases: an epidemiological perspective. *Philos. Trans. R. Soc. B* 363(1492):741–59
48. Glais L, Kerlan C, Tribodet M, Marie-Jeanne Tordo V, Robaglia C, Astier-Manifacier S. 1996. Molecular characterization of potato virus Y^N isolates by PCR-RFLP. *Eur. J. Plant Pathol.* 102(7):655–62
49. Goss EM. 2015. Genome-enabled analysis of plant-pathogen migration. *Annu. Rev. Phytopathol.* 53:121–35
50. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, et al. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656):327–32
51. Grünwald NJ, Goss EM. 2011. Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches. *Annu. Rev. Phytopathol.* 49:249–67
52. Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Mol. Ecol.* 18(23):4734–56
53. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. 2013. Robust estimation of microbial diversity in theory and in practice. *ISME J.* 7(6):1092–101
54. Hall M, Woolhouse M, Rambaut A. 2015. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLOS Comput. Biol.* 11(12):e1004613
55. Hampson K, Dushoff J, Cleaveland S, Haydon DT, Kaare M, et al. 2009. Transmission dynamics and prospects for the elimination of canine rabies. *PLOS Biol.* 7(3):e1000053
56. Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, et al. 2003. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. R. Soc. B* 270(1511):121–27
57. Heller R, Chikhi L, Siegismund HR. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLOS ONE* 8(5):e62992
58. Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* 23(24):5947–65
59. Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* 11(3):423–34
60. Holmes EC. 2009. The evolutionary genetics of emerging viruses. *Annu. Rev. Ecol. Evol. Syst.* 40:353–72
61. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67(10):4399–406
62. Jeger MJ, Seal SE, Van den Bosch F. 2006. Evolutionary epidemiology of plant virus disease. *Adv. Virus Res.* 67:163–203

This review sents the principles he phylodynamics nework, which ates phylogenies and demiology.

This review sents the different lecular clock models.

64. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLOS Comput. Biol.* 10(1):e1003457
65. Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 11:94
66. Jombart T, Devillard S, Dufour A-B, Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* 101(1):92–103
67. Jombart T, Eggo RM, Dodd PJ, Balloux F. 2011. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 106(2):383–90
68. Jridi C, Martin J-F, Marie-Jeanne V, Labonne G, Blanc S. 2006. Distinct viral populations differentiate and evolve independently in a single perennial host plant. *J. Virol.* 80(5):2349–57
69. Kamvar ZN, Brooks JC, Grünwald NJ. 2015. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* 6:208
70. Kenah E, Britton T, Halloran ME, Longini IM Jr. 2016. Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. *PLOS Comput. Biol.* 12(4):e1004869
71. Klinkowski M. 1970. Catastrophic plant diseases. *Annu. Rev. Phytopathol.* 8:37–60
72. Kraberger S, Harkins GW, Kumari SG, Thomas JE, Schwinghamer MW, et al. 2013. Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* 444(1–2):282–91
73. Lacroix C, Renner K, Cole E, Seabloom EW, Borer ET, Malmstrom CM. 2016. Methodological guidelines for accurate detection of viruses in wild plant species. *Appl. Environ. Microbiol.* 82(6):1966–75
74. Lau MSY, Marion G, Streftaris G, Gibson G. 2015. A systematic Bayesian integration of epidemiological and genetic data. *PLOS Comput. Biol.* 11(11):e1004633
75. Lefeuvre P, Martin DP, Harkins G, Lemey P, Gray AJA, et al. 2010. The spread of *Tomato yellow leaf curl virus* from the Middle East to the world. *PLOS Pathog*. 6(10):e1001164
76. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, et al. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLOS Pathog*. 10(2):e1003932
77. Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* 5(9):e1000520
78. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* 27(8):1877–85
79. Logan G, Freimanis GL, King DJ, Valdazo-González B, Bachanek-Bankowska K, et al. 2014. A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC Genom*. 15:828
80. Mabvakure B, Martin DP, Kraberger S, Cloete L, van Brunschot S, et al. 2016. Ongoing geographical spread of *Tomato yellow leaf curl virus*. *Virology* 498:257–64
81. Maio ND, Wu CH, O'Reilly KM, Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLOS Genet*. 11(8):e1005421
82. Malmstrom CM, Shu R, Linton EW, Newton LA, Cook MA. 2007. Barley yellow dwarf viruses (BYDVS) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. *J. Ecol.* 95(6):1153–66
83. Marais A, Faure C, Couture C, Bergey B, Gentit P, Candresse T. 2014. Characterization by deep sequencing of divergent *Plum bark necrosis stem pitting–associated virus* (PBNSPaV) isolates and development of a broad-spectrum PBNSPaV detection assay. *Phytopathology* 104(6):660–66
84. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. 2011. Recombination in eukaryotic single stranded DNA viruses. *Viruses* 3(9):1699–738
85. Martin DP, Lemey P, Posada D. 2011. Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.* 11(6):943–55
86. McRae BH. 2006. Isolation by resistance. *Evolution* 60(8):1551–61

67. This study presents the first R package enabling to reconstruct transmission graphs from genetic data.

77. This article presents a Bayesian framework for inference, visualization, and hypothesis testing of phylogeographic history.

81. This paper introduces BASTA, a phylogeographic model combining accurate coalescence methods and computational efficiency.

This article scribes a framework reconstructing nsmission trees using netic, temporal, and tial information.

88. **Mollentze N, Nel LH, Townsend S, Roux K, Hampson K, et al. 2014. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B* 281(1782):20133251**
89. Monjane AL, Harkins GW, Martin DP, Lemey P, Lefeuvre P, et al. 2011. Reconstructing the history of Maize streak virus strain A dispersal to reveal diversification hot spots and its origin in southern Africa. *J. Virol.* 85(18):9623–36
90. Monmonier MS. 1973. Maximum-difference barriers: an alternative numerical regionalization method. *Geogr. Anal.* 5(3):245–61
91. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. 2012. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLOS Comput. Biol.* 8(11):e1002768
92. Moury B, Desbiez C, Jacquemond M, Lecoq H. 2006. Genetic diversity of plant virus populations: towards hypothesis testing in molecular epidemiology. *Adv. Virus Res.* 67:49–87
93. Moury B, Fabre F, Senoussi R. 2007. Estimation of the number of virus particles transmitted by an insect vector. *PNAS* 104(45):17891–96
94. Mullan B, Sheehy P, Shanahan F, Fanning L. 2004. Do Taq-generated RT-PCR products from RNA viruses accurately reflect viral genetic heterogeneity? *J. Viral Hepat.* 11(2):108–14
95. Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, et al. 2016. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* 7(1):80–89
96. Nelson MI, Viboud C, Vincent AL, Culhane MR, Detmer SE, et al. 2015. Global migration of influenza A viruses in swine. *Nat. Commun.* 6:6696
97. Ng TFF, Duffy S, Polston JE, Bixby E, Vallad GE, Breitbart M. 2011. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLOS ONE* 6(4):e19050
98. O'Fallon BD. 2013. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinform.* 14:40
99. Ohshima K, Akaishi S, Kajiyama H, Koga R, Gibbs AJ. 2010. Evolutionary trajectory of turnip mosaic virus populations adapting to a new host. *J. Gen. Virol.* 91(3):788–801
100. Ohshima K, Matsumoto K, Yasaka R, Nishiyama M, Soejima K, et al. 2016. Temporal analysis of reassortment and molecular evolution of *Cucumber mosaic virus*: extra clues from its segmented genome. *Virology* 487:188–97
101. Olarte Castillo XA, Fermin G, Tabima J, Rojas Y, Tennant PF, et al. 2011. Phylogeography and molecular epidemiology of *Papaya ringspot virus*. *Virus Res.* 159(2):132–40
102. Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *PNAS* 86(8):2766–70
103. Ostfeld RS, Glass GE, Keesing F. 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends Ecol. Evol.* 20(6):328–36
104. Palanga E, Filloux D, Martin DP, Fernandez E, Gargani D, et al. 2016. Metagenomic-based screening and molecular characterization of cowpea-infecting viruses in Burkina Faso. *PLOS ONE* 11(10):e0165188
105. Parker IM, Gilbert GS. 2004. The evolutionary ecology of novel plant-pathogen interactions. *Annu. Rev. Ecol. Evol. Syst.* 35:675–700
106. Piry S, Chapuis M-P, Gauffre B, Papaïx J, Cruaud A, Berthier K. 2016. Mapping averaged pairwise information (MAPI): a new exploratory tool to uncover spatial structure. *Methods Ecol. Evol.* 7:1463–75
107. Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *PNAS* 98(24):13757–62
108. Prasanna H, Sinha DP, Verma A, Singh M, Singh B, et al. 2010. The population genomics of begomoviruses: global scale population structure and gene flow. *Virol. J.* 7:220
109. Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10(8):540–50
110. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, et al. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *PNAS* 109(37):15066–71

112. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530(7589):228–32
113. R Dev. Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Found. Stat. Comput.
114. Rakotomalala M, Pinel-Galzi A, Mpunami A, Randrianasolo A, Ramavovololona P, et al. 2013. *Rice yellow mottle virus* in Madagascar and in the Zanzibar Archipelago; island systems and evolutionary time scale to study virus emergence. *Virus Res*. 171(1):71–79
115. Rasmussen DA, Ratmann O, Koelle K. 2011. Inference for nonlinear epidemiological models using genealogies and time series. *PLOS Comput. Biol.* 7(8):e1002136
116. Rees EE, Pond BA, Cullingham CI, Tinline R, Ball D, et al. 2008. Assessing a landscape barrier using genetic simulation modelling: implications for raccoon rabies management. *Prev. Vet. Med.* 86(1–2):107–23
117. Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol. Ecol.* 25(9):1911–24
118. Rissler LJ. 2016. Union of phylogeography and landscape genetics. *PNAS* 113(29):8079–86
119. Rodríguez-Cerezo E, Moya A, García-Arenal F. 1989. Variability and evolution of the plant RNA virus pepper mild mottle virus. *J. Virol.* 63(5):2198–203
120. Rolland M, Glais L, Kerlan C, Jacquot E. 2008. A multiple single nucleotide polymorphisms interrogation assay for reliable Potato virus Y group and variant characterization. *J. Virol. Methods* 147(1):108–17
121. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T. 2014. Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Mol. Biol. Evol.* 31(9):2472–82
122. Roossinck MJ. 2014. Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Virology* 5:767
123. Roossinck MJ, Martin DP, Roumagnac P. 2015. Plant virus metagenomics: advances in virus discovery. *Phytopathology* 105(6):716–27
124. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. 1977. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 265:687–95
125. Sanjuán R. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. *PLOS Pathog*. 8(5):e1002685
126. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J. Virol.* 84(19):9733–48
127. Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156(2):879–91
128. Schubert J, Habekuß A, Kazmaier K, Jeske H. 2007. Surveying cereal-infecting geminiviruses in Germany—diagnostics and direct sequencing using rolling circle amplification. *Virus Res.* 127(1):61–70
129. Schulte PA, Perera FP. 1993. *Molecular Epidemiology: Principles and Practices*. San Diego: Academic
130. Seo TK, Thorne JL, Hasegawa M, Kishino H. 2002. A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* 18(1):115–23
131. Shah VB, McRae BH. 2008. Circuitscape: a tool for landscape ecology. *Proc. Python Sci. Conf.*, 7*th*, *Pasadena*, Aug. 19–24, pp. 62–65. **https://hal.archives-ouvertes.fr/hal-00502586**
132. Shapiro B, Ho SYW, Drummond AJ, Suchard MA, Pybus OG, Rambaut A. 2011. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* 28(2):879–87
133. Simmons HE, Dunham JP, Stack JC, Dickins BJA, Pagán I, et al. 2012. Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. *J. Gen. Virol.* 93(8):1831–40
134. Smith O, Clapham A, Rose P, Liu Y, Wang J, Allaby RG. 2014. A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep*. 4:4003
135. Soubeyrand S. 2016. Construction of semi-Markov genetic-space-time SEIR models and inference. *J. Soc. Fr. Stat.* 157(1):129–52

117. This review summarizes tip dating approaches and provides a guide to performing such analyses.

137. Stainton D, Martin DP, Muhire BM, Lolohea S, Halafihi M, et al. 2015. The global distribution of *Banana bunchy top virus* reveals little evidence for frequent recent, human-mediated long distance dispersal events. *Virus Evol.* 1(1):vev009
138. Sztuba-Solińska J, Urbanowicz A, Figlerowicz M, Bujarski JJ. 2011. RNA-RNA recombination in plant virus replication and evolution. *Annu. Rev. Phytopathol.* 49:415–43
139. Tomimura K, Špak J, Katis N, Jenner CE, Walsh JA, et al. 2004. Comparisons of the genetic structure of populations of *Turnip mosaic virus* in West and East Eurasia. *Virology* 330(2):408–23
140. Trovão NS, Baele G, Vrancken B, Bielejec F, Suchard MA, et al. 2015. Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* 1(1):vev016
141. van Etten J. 2015. *R Package Gdistance: Distances and Routes on Geographical Grids*. **https://cran.r-project.org/package = gdistance**
142. Viboud C, Nelson MI, Tan Y, Holmes EC. 2013. Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. *Philos. Trans. R. Soc. B* 368(1614):20120199
143. Vurro M, Bonciani B, Vannacci G. 2010. Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences. *Food Secur.* 2(2):113–32
144. Wang IJ. 2010. Recognizing the temporal distinctions between landscape genetics and phylogeography. *Mol. Ecol.* 19(13):2605–8
145. Wheeler DC, Waller LA, Biek R. 2010. Spatial analysis of feline immunodeficiency virus infection in cougars. *Spat. Spatiotemporal Epidemiol.* 1(2–3):151–61
146. Worby CJ, Lipsitch M, Hanage WP. 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLOS Comput. Biol.* 10(3):e1003549
147. Yasaka R, Nguyen HD, Ho SYW, Duchêne S, Korkmaz S, et al. 2014. The temporal evolution and global spread of *Cauliflower mosaic virus*, a plant pararetrovirus. *PLOS ONE* 9(1):e85641
148. Yasaka R, Ohba K, Schwinghamer MW, Fletcher J, Ochoa-Corona FM, et al. 2015. Phylodynamic evidence of the migration of turnip mosaic potyvirus from Europe to Australia and New Zealand. *J. Gen. Virol.* 96(3):701–13
149. Ypma RJF, van Ballegooijen WM, Wallinga J. 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195(3):1055–62
150. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. 2012. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B* 279(1728):444–50

Résultats clés de l'Article 2

EXPLOITER L'INFORMATION GENETIQUE POUR EVALUER LA DISPERSION DES VIRUS DANS LES PAYSAGES

Cette synthèse bibliographique analyse les approches d'épidémiologie moléculaire qui ont été développées pour suivre la dispersion des virus dans les paysages. Les principaux points à retenir sont les suivants :

1. Le taux de substitution élevé des virus implique que les processus évolutifs et épidémiologiques sont observables à la même échelle de temps. Les génomes viraux peuvent donc être utilisés pour inférer la dynamique du virus dans les paysages à travers le temps et l'espace.

2. Les progrès des méthodes de caractérisation des virus ont permis d'élargir nos connaissances sur les espèces virales existantes et sur leur diversité intraspécifique.

3. Des plans d'échantillonnage appropriés sont nécessaires pour éviter les biais lorsque l'on étudie l'influence des variables biologiques et environnementales sur la structuration de la diversité des populations virales.

4. Les virus étant clonaux, les analyses exploratoires sans hypothèse sont plus appropriées que les approches classiques de génétique des populations pour décrire la structure spatiale de la diversité virale.

5. Les modèles phylogéographiques permettent d'inférer des voies d'invasion sur de grandes zones à partir de coordonnées géographiques de séquences datées.

6. De nouvelles approches combinant la génétique à l'échelle du paysage (« landscape genetics ») et la phylogéographie permettent de tester l'impact de la configuration et de la composition du paysage sur la dynamique spatio-temporelle du virus.

7. De récents modèles phylodynamiques et des modèles SEIR prenant en compte des données épidémiologiques et génétiques peuvent être utilisés pour inférer des arbres de transmission et des paramètres épidémiologiques clés (en se basant sur des séquences de virus provenant d'épidémies intensément échantillonnées).

2. Application au virus de la sharka

Afin d'estimer les paramètres épidémiologiques d'une maladie à l'échelle de l'hôte, un modèle généticospatio-temporel a été développé dans l'unité BioSP à Avignon, en collaboration avec des chercheurs de BGPI et de l'université de Glasgow (Mollentze et al. 2014, Morelli et al. 2012, Soubeyrand, 2016). Ce modèle prend en compte des données épidémiologiques et génétiques, et permet d'inférer « qui a infecté qui » dans un paysage, ainsi que des paramètres épidémiques clés.

Dans un premier temps, nous avons testé l'efficacité de ce modèle sur des données simulées. Pour cela, j'ai réalisé des simulations d'épidémies de sharka à l'aide du modèle développé par Pleydell et al. (2018) et Rimbaud et al. (2018a, 2018b) que j'ai adapté pour obtenir des données à l'échelle de l'arbre. Ces simulations ont permis de comparer par simulation la précision de la reconstruction des chaines de transmission entre les hôtes avec ou sans la prise en compte des arbres non infectés. Ce test était important pour ensuite appliquer le modèle généticospatio-temporel à des données réelles de manière efficace.

Par la suite, nous avons tenté d'estimer certains paramètres épidémiologiques de la sharka en appliquant le modèle de reconstruction des chaines de transmission sur des données réelles. Le PPV est un virus à ARN dont le génome (10 kb environ) évolue rapidement. Il nous est donc paru possible de reconstruire les chaines de transmission entre les hôtes et d'estimer les paramètres épidémiologiques sous-jacents. L'application de cette approche à la sharka n'a cependant pas été finalisée car des adaptations de la méthode sont vraisemblablement encore nécessaires pour fournir des résultats robustes. Ces travaux sont présentés dans l'article 3 pour lequel j'ai contribué à l'écriture et à la production des résultats des parties qui traitent des simulations des données de dispersion d'une épidémie et des données réelles de sharka. J'ai également participé à la mise en forme des données qui ont servi à réaliser l'inférence des chaines de transmission ainsi qu'à l'analyse des résultats.

ARTICLE 3

Accounting for uninfected hosts in transmission tree reconstruction

Coralie Picard, Sylvie Dallot, Gaël Thébaud and Samuel Soubeyrand

Accounting for uninfected hosts in transmission tree reconstruction

Picard C. (1), Dallot S. (1), Thébaud G. (1) and Soubeyrand S. (2)

(1) BGPI, INRA, Montpellier SupAgro, Univ. Montpellier, Cirad, 34398, Montpellier, France

(2) BioSP, INRA, 84914, Avignon, France.

ABSTRACT

Several approaches coupling epidemiological and evolutionary models, genetic-space-time data and appropriate inference techniques have been proposed to infer transmissions in outbreaks. These approaches are grounded on data, which generally do not contain information on hosts that are not infected during the observation period. The absence of negative data is generally caused by the large number of uninfected hosts compared to the number of infected hosts in studies where the approaches for inferring transmissions were tested. Here, we precisely study the impact of including uninfected hosts in the inference of transmissions in the context of plant epidemiology. For that purpose, we modified an existing genetic-space-time approach allowing the estimation of "who infected whom" by incorporating uninfected hosts in the underlying epidemiological model, and we assessed the advantage of incorporating such hosts in a numerical study based on simulated outbreaks of a plant pathogen (*Plum pox virus*). We showed that integration of uninfected hosts allowed reconstructing 35% of the transmissions (against 20% without it). Thus, including uninfected hosts in a joint analysis of epidemiological and genetic data provides a better understanding of the spatial epidemiology of a pathogen and provides valuable insights into transmission dynamics. Such knowledge on transmissions is crucial for designing efficient control policies.

Keywords: transmission tree, space time genetic, sharka, landscape, SEIR

1. Introduction

Epidemics caused by pathogen spread through host populations can be a high socioeconomic burden (Klinkowski, 1970; Vurro et al., 2010). In order to minimize the associated costs, governmental agencies often design management strategies relying on scientific expertise. To support public policy decision-making, scientists need to understand and predict how pathogens spread within and between host populations (Ferguson et al., 2003; Keeling et al., 2003). More specifically, the reconstruction of transmission routes during past epidemics may help to predict how the same pathogens will spread through similar populations in future outbreaks (Picard et al., 2017). Indeed, understanding the history of transmission events can highlight key drivers of transmission, provide refined estimates of epidemiological parameters and point out risk factors related to vectors, reservoirs and landscape components, which can help build epidemiological projections (Ostfeld et al., 2005).

However, identifying transmission links between hosts in a landscape remains a challenging task. Indeed, the locations of diseased individuals through time are usually consistent with many different transmission trees (i.e. "who-infected-who"). Pathogen genome sequences collected during epidemics can help discriminating between such trees, because genetic data can provide critical additional information regarding the relationships between hosts infected by measurably evolving pathogens (i.e. that fix mutations across their genome during the course of a single outbreak; Picard et al., 2017).

Various models integrating genetic and spatiotemporal data have been developed to understand transmission links between hosts (Jombart et al., 2014; Lau et al., 2015; Mollentze et al., 2014; Morelli et al., 2012; Ypma et al., 2012, 2013; Worby et al., 2014). These models enable to infer epidemiological processes, specifically the most likely transmission tree reflecting "who infected whom", and other parameters related to the infection strength, the latent period, the incubation period, the dispersal kernel (which partly determines the speed and spatial extent of disease spread), and the substitution rate (Soubeyrand, 2016). Such genetic-space-time models are generally stochastic and based on an SEIR (Susceptible, Exposed, Infectious, Removed) structure explicitly representing successive sanitary statuses of host individuals.

For now, these models have been used for animal and human diseases but, to our knowledge, they have never been applied to plant diseases. In addition, they never accounted for the localization of the uninfected hosts, since such data can be difficult to obtain, particularly for animal and human diseases. Regarding plant diseases, the locations of uninfected hosts are more easily available and

have already been used to infer disease transmissions in the absence of genetic data (Gibson, 1997; Neri et al., 2014). However, the number of uninfected hosts is generally much higher than the number of infected hosts; thus, it can become challenging to account for them in genetic-space-time models. Indeed, a key challenge in plant disease modelling is to assess the impact of incomplete host data on model predictions (Cunniffe et al., 2015).

In this context, we aimed to understand how taking into account uninfected hosts into transmission tree reconstruction can improve the estimation of "who infected whom". For that purpose, we modified an existing genetic-space-time SEIR model and its associated estimation method (Mollentze et al., 2014; Soubeyrand, 2016). Then, we assessed inference performance using simulated data obtained by coupling a micro-evolutionary model of pathogen sequences with a spatio-temporal epidemiological model built for sharka (Picard et al., in prep; Picard et al., in revision; Pleydell et al., 2018; Rimbaud et al., 2018). Sharka is one of the most damaging diseases of stone fruit trees belonging to the genus *Prunus* (e.g. peach, apricot and plum) (Cambra et al., 2006; Rimbaud et al., 2015) and it is caused by *Plum pox virus* (PPV, *Potyviru*s genus). As many RNA viruses, PPV is expected to evolve quickly and its evolutionary and epidemic dynamics are supposed to happen at similar time scales.

2. Materials and methods

2.1. Genetic-space-time SEIR model

In this article, we extended the genetic-space-time SEIR model described by Soubeyrand (2016) representing the transmissions of an infectious disease within a population of susceptible hosts and the micro-evolution of the pathogen causing the disease. The genetic-space-time SEIR model results from the coupling of a semi-Markov, individual-based, continuous-time, spatial epidemic model that governs the transitions between the sanitary statuses of individuals (S, E, I and R) and a Markovian evolutionary model that governs nucleotide substitutions in the sequence of the pathogen at the host level. This model was extended to apply it to sharka epidemics: it handles (i) the emergence of hosts across the study period, (ii) the delay between the detection of infected hosts and their removal.

Tables 1 and 2 describe the epidemiological and evolutionary events that are included in the genetic-space-time SEIR model. Mathematical details are provided in Soubeyrand (2016). Here, we simply

comment on the points related with the model extension. First, the delay between the detection of infected hosts and their removal is treated like in Morelli et al. (2012). The second extension relates to host emergence (here, host plantation), which is supposed to occur at known dates and in the healthy state. If a host is actually infected before plantation, the method for inferring transmissions should select early dates of infection for this host, which would mimic an infection at the plantation date.

Table 1. Possible events and corresponding transition rates or distributions for the semi-Markov, individual-based, continuous-time, spatial epidemic model SEIR model. Host *k* emerges at its (known) date of plantation. Then, after infection, it enters the exposed stage at the rate given in the table; this rate is defined as the sum of a basic risk α_0 and the contributions of infectious hosts at time *t* weighted by a kernel *w* computed at the distances d_{jk} between the focal host *k* and the infectious hosts. *w* was specified as the 2D exponential kernel parameterized by γ: $w(d) = \exp\left(-\frac{d}{\gamma}\right)/(2\pi\gamma^2)$. The duration of the exposed stage (latency period) and the duration between the end of the exposed stage and the detection of the infected host are drawn from a gamma distribution. Note that the gamma distribution is parameterized here by its mean and its standard deviation. Finally, host *k* is removed at the (known) uprooting date.

Description	**Event**	**Rate**	**Distribution**[a]
Host emergence	S_k: 0 → 1		Dirac(plantation date)
Infection	S_k: 1 → 0 & E_k: 0 → 1	$\alpha_0 + \alpha_1 \sum_{j \neq k} w(d_{jk}) I_j(t)$	
Beginning of infectious stage	E_k: 1 → 0 & I_k: 0 → 1		Gamma(β_1 , β_2)
Detection	I_k: 1 → 1		Gamma(δ_1 , δ_2)
End of infectious stage	I_k: 1 → 0 & R_k: 0 → 1		Dirac(uprooting date)

[a] For Infection, the equation corresponds to a rate.

Table 2. Possible events and corresponding substitution rates for the Markovian evolutionary model. Letters A, C, G and U denotes nucleotides adenine, cytosine, guanine and uracil, respectively.

Description	**Event**	**Rate**
Transition	A→G or G→A or C→U or U→C	μ_1
Transversion (type 1)	A→U or U→A or C→G or G→C	μ_2
Transversion (type 2)	A→C or C→A or G→U or U→G	μ_3

2.2. Estimation method

The estimation of model parameters and latent variables (sources of infection, infection times and durations of exposed stages) was carried out in the Bayesian framework, following the method of Soubeyrand (2016) based on the approximate genetic likelihood. This method includes the reconstruction of sequences that are transmitted at the infection events by using a parsimonious reconstruction algorithm. Only direct transmissions were reconstructed (indirect transmissions handled in Jombart et al. (2014) and Mollentze et al. (2014) were not taken into account here). For each treated dataset, the posterior distribution was evaluated with three interacting MCMC chains (chain length: 10^5; burn-in: 4000 iterations; thinning: every 100 iterations; interaction between chains: every 2000 iterations).

Two versions of the estimation method were run for each dataset: hosts that remained healthy were either included in or removed from the dataset. Incorporating healthy hosts in the estimation method amounts to compute for these hosts the probability that they have not been infected up to the end of the observation period. This probability is incorporated into the transmission likelihood (see Soubeyrand, 2016). To handle healthy hosts, we ignored the possibility that apparently healthy hosts were actually infected.

2.3. Simulated data

2.3.1 Simulation of sharka epidemics

In order to assess the performance of the estimation method (with and without the uninfected hosts), we used an existing simulation model of sharka disease (Picard et al., in prep; Picard et al., in revision; Pleydell et al., 2018; Rimbaud et al., 2018). This stochastic, spatially explicit SEIR model includes 6 epidemiological parameters characterizing the epidemic, and 21 disease management parameters enabling to simulate orchard surveillance, plantation bans and removal of infected trees. Here, we used the same variation ranges of epidemiological parameters as in by Picard et al. (in prep) to simulate 20 established epidemics. In addition, for each simulation, the epidemic spreads during 5 years without management, followed by 10 years with one survey per year performed with a detection probability of 0.66 (once an infected tree is detected, it is removed from the simulation).

These epidemics were run on a virtual landscape comprising 2508 trees grouped into 16 patches (Fig 1). To simulate this landscape, we generated a grid pattern of 10,000 squares (16 m² each) representing potential trees. Among them, we randomly selected 20 rectangular patches with

random width and length (between 2 and 21 trees). When two rectangular patches included the same trees, we grouped them into a single patch (hence the 16 final patches).

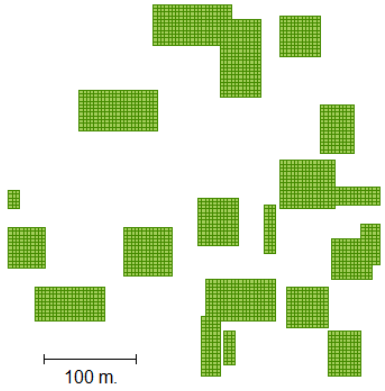


Figure 1: Simulated landscape composed of 2508 trees allocated into 16 different patches

2.3.2. Simulation of pathogen sequence evolution

The evolution of PPV in the hosts was simulated conditional on the transmission tree obtained from the simulation of the epidemics. We used sequence fragments of 10,000 nucleotides, among which a fraction was fixed. We assumed that there is no within-host diversity, i.e. at any time each host is infected by at most one genomic sequence. When a tree was infected at plantation, the sequence of the pathogen was drawn from a set of 20 slightly varying reference sequences (each of these 20 sequences were obtained by uniformly randomly modifying any nucleotide with a 5‰ chance from a reference genome). When a tree was infected by another one within the simulated landscape, the current PPV sequence infecting the source tree was used as the initial sequence in the receiving host. The substitution of nucleotides within each host was performed forward in time as a Markov chain with heterogeneous rates of substitution across the sequence. More specifically, the rate of substitution for each nucleotide was drawn from a zero-inflated gamma distribution with the probability of zero equal to 0.3 (the shape of the gamma distribution was 0.3 and its scale parameter was 10^{-4}). This substitution rate (10^{-4} subs/site/year) was estimated under a Bayesian framework (BEAST 1.8 software) using 86 heterochronous whole genome sequences of PPV isolates sampled in peach orchards of southern France from 1991 to 2008 (Dallot et al., 2016).

2.3.3. Subsampling for generating datasets used in the inference

We simulated 20 outbreaks through the 16 patches and we retrieved the geographical coordinates of the trees, as well as their plantation and removal dates. We also got the simulated dates of disease surveillance, the sanitary status of the trees at each date (symptomatic/ non-symptomatic) as well as the simulated viral sequences associated with the infected trees. In order to reduce the overall computational cost of outbreak reconstructions, we did not attempt to reconstruct the transmissions between all the trees in the landscape simultaneously, but only between trees of some patches (note that this approach reflects a frequent situation since epidemiological and genetic data are generally available for only a part of the landscape). For each outbreak, we selected the patches with the highest number of infected prunus trees (excluding patches with more than 400 trees), without exceeding a total of 1200 trees (healthy and infected) on all patches. The data corresponding to these patches were used to reconstruct transmission chains between hosts with the estimation method introduced above, both with and without the non-infected trees.

2.4. Specification of prior distributions for the inference

For the genetic-space-time SEIR model, vague exponential priors with mean 100 were used for the infection strengths α_0 and α_1, corresponding respectively to exogenous (from trees not included in the dataset) and endogenous (from trees included in the dataset) sources. In addition, informative gamma priors with mean and standard deviation equal to 331 m was specified for the mean dispersal distance 2γ (331 m was the mean dispersal distance derived from the estimation for sharka in Pleydell et al. (2018), to 1.92 yr and 0.1 for the mean incubation duration θ_1, to 0.66 yr and 0.1 for the standard deviation of the incubation duration θ_2 (these prior means were taken from Pleydell et al., 2018), to 3 yr and 0.1 for the mean duration δ_1 between the end of the exposed stage and the detection, and to 2 yr and 0.1 for the standard deviation δ_2. Vague exponential priors with mean 10^{-4} sub/site/year were used for the substitution rates μ_1, μ_2 and μ_3 (Dallot et al., 2016). From year 1, the first year of the epidemic, a normal prior with mean -100 yr and standard deviation 50 yr was used for the time of the most recent common ancestor.

3. Results

We tried to reconstruct the transmission trees with a median of 45 infected trees for each of the 20 simulations performed. Among these infected trees, on average 20 trees were infected by an external source (i. e. by a tree located outside the simulated landscape, and for which we do not have informations), and on average 26 by a tree located in the landscape, for which we have epidemiological and genetic data (Fig 2A and B).

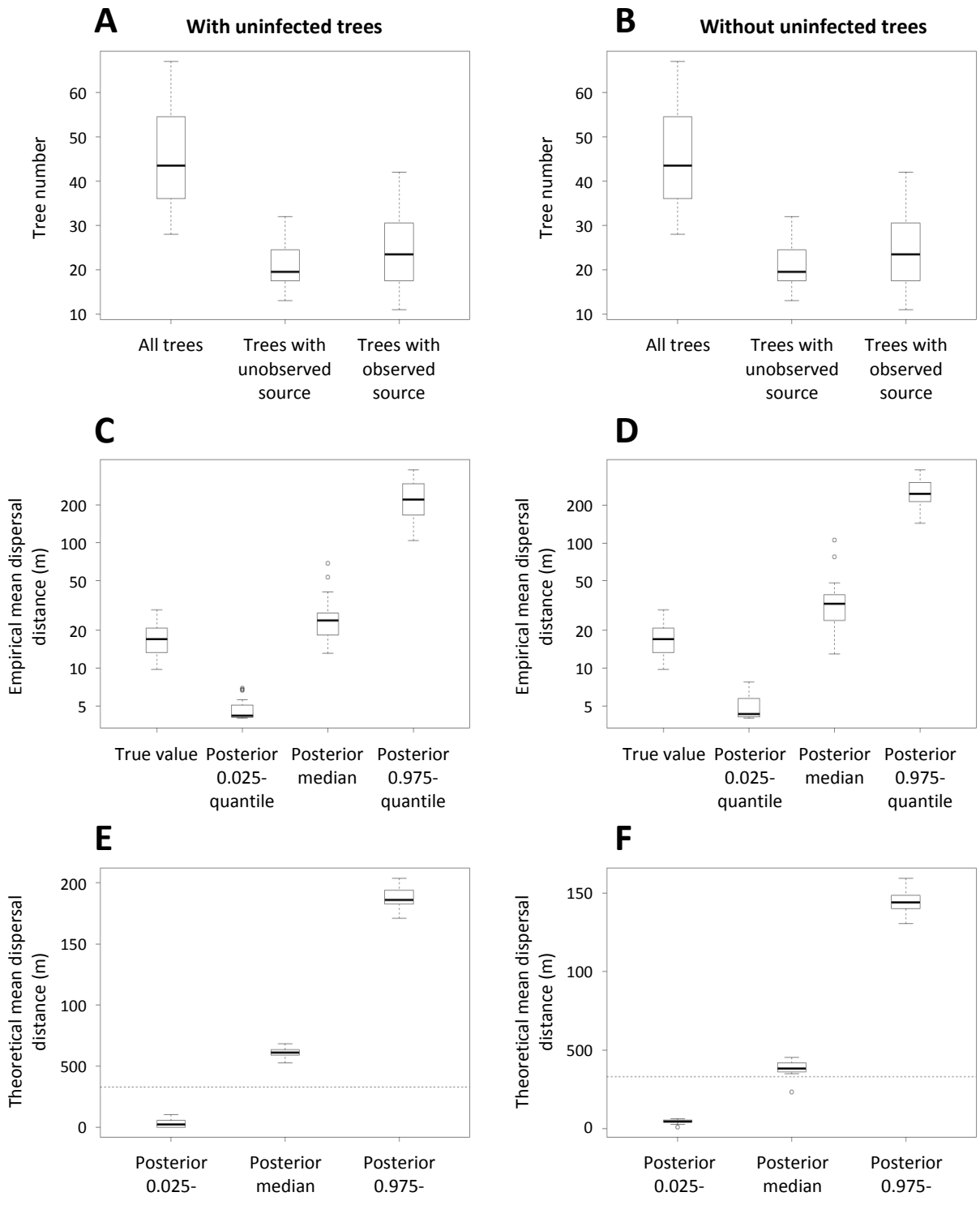


Figure 2: Results of the inference of simulated sharka epidemics. A and B: among the 20 simulated outbreaks, frequency of all infected trees, and frequency of infected trees whose source of infection is located outside (unobserved source) or inside (observed source) the selected orchards. C and D: empirical mean dispersal distance (calculated from the transmissions between trees located in the simulated landscape). E and F: Theoretical mean dispersal distance (accounting for the external transmissions). A, C and E: results with uninfected trees. B, D and F: results without uninfected trees.

The genetic-space-time model allowed us to reconstruct 35% of the transmissions when the uninfected trees were considered but only 20% without taking them into account (Fig 3). As a comparison, we would have reconstructed only 0.02% of the transmissions by randomly generating transmissions trees. In addition, we identified less transmissions when the source trees were located in the landscape (30% with the uninfected trees and 19% without), than when the trees were infected by an external source (45% with the uninfected trees and 21% without).

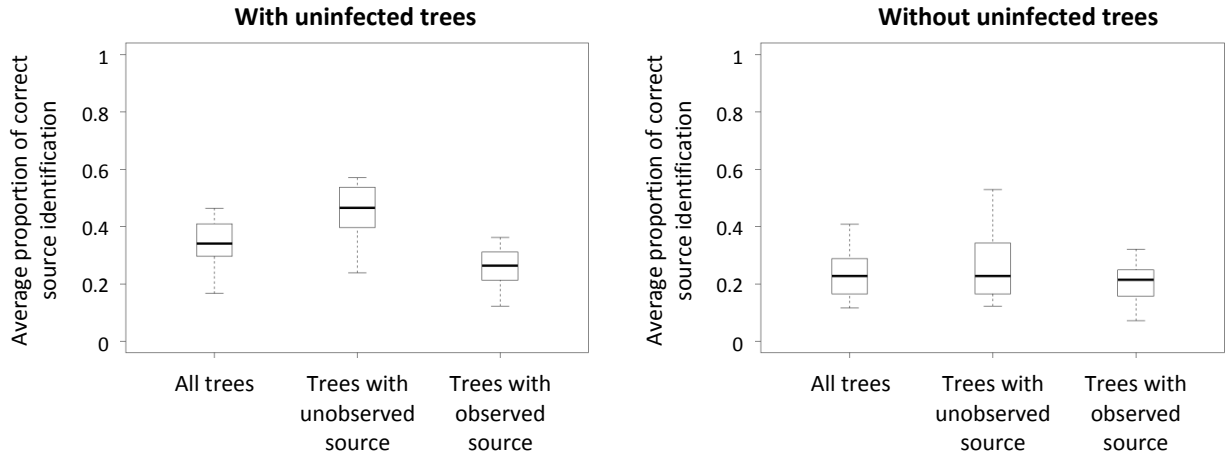


Figure 3: Accuracy of the reconstruction of transmissions assessed from simulated sharka epidemics.

The mean of the empirical dispersal (calculated from the transmissions between infected prunus trees located in the simulated landscape) was correctly estimated when taking into account the uninfected trees. Indeed, the median of the empirical dispersal was estimated at 23m (accounting for uninfected trees), and the true value (calculated from the simulations) at 18m (Fig 2C and D). Without the uninfected trees, the median of the empirical dispersal was slightly overestimated (30m). However, in both cases, the estimated quantile distributions were far from the true value of the empirical dispersal.

We also estimated the prediction accuracy of the theoretical dispersal, which differs from the empirical dispersal accounting for the external transmissions in its calculation. The estimation of the mean of theoretical dispersal was overestimated with or without the uninfected trees (although the prediction accuracy was higher without the uninfected trees). Indeed, in the sharka simulation model, the mean of theoretical dispersal was 331m, and the estimations were 600 and 400 with and without the uninfected trees respectively (Fig 2E and F). However, it was difficult to estimate this

epidemiological parameter since we performed our simulations in a small area (only slightly bigger than 331m).

4. Discussion

4.1. Conclusion and applicability of the approach

In this study, we showed how including uninfected hosts in a genetic-space-time model can improve the reconstruction of transmission trees. For that purpose, we used the example of sharka disease, for which we simulated dispersal and management, as well as the genetic sequence of the virus for each infected tree. Then, the genetic-space-time model allowed us to reconstruct the transmission links from these simulated data. We showed that accounting for the uninfected trees improved the inference of transmission links: we reconstructed 35% of the transmissions with the uninfected trees (against 20% without).

However, the epidemics were here simulated through landscapes for which the orchards are composed of few trees (the bigger simulated orchard includes only 231 trees), which represents an area with traditional arboriculture. By contrast, more recent exploitations are generally composed of bigger orchards (which can include more than 1000 trees). In such situation, the performance of the transmission tree reconstruction may be different since the landscape present less discontinuity between hosts. However, testing the genetic-space-time model in this case can increase the inference duration (which is multiplied increasing the number of trees).

Our results could allow improving numerous studies which aim to understand and to predict how pathogens spread within host populations, which could help to develop adapted management strategies to control pathogens. This approach is particularly interesting for diseases of perennial plants since they are localized at the same place during several years. However, using genetic-space-time model on diseases which spread on annual hosts can be more challenging since it is difficult, if not impossible, to follow the temporal signal included in the genetic sequences (which is essential to perform the inference). Similarly, our approach could be difficult to transpose to human and animal diseases since the hosts are generally mobile. To address this issue, it could be interesting to account for the host movements in the inference, but aside the need of lot of material, tracking them can cause ethical problems.

4.2. Transmission links inference of real sharka data

To go even further in this study, we attempted to use the genetic-space-time model on a real sharka epidemic. The material and method used is described in S1 text. However, the distribution of the transmissions between trees was unsatisfactory. Indeed, transmission links obtained with the inference were characterized by a small number of trees which infected numerous other trees located further in the landscape. We would have expected that the model infer less long distance transmissions and more local transmissions (short distance). These unsatisfactory results are probably due to a lack of information from the dataset and a maladaptation of the method for these data. Indeed, we attempted to reconstruct the transmissions trees with data only sampled over 3 consecutive years, the temporal signal was thus difficult to capture, especially because latency duration of sharka may vary from few weeks to few years.

In order to improve this preliminary work, we tried to remove from the dataset the infected trees for which we did not know the virus genetic sequence. Indeed, the inference suggested that these trees were the source of most of the transmissions, which is unlikely. In addition, we modified the prior of the model corresponding to the mutation rate in order to limit the possibility of long transmissions. Thus, we inferred transmissions links between infected trees with a transmission rate of 10^{-5} subs/site/year instead of 10^{-4} subs/site/year. Nevertheless, these two attempts to improve the inference were not satisfactory. However, we did not explore ways that could improve the estimation of transmissions. Firstly, although the inference can be much longer, we could perform it with all our available data (for now, we carried out the inference on only 6 out of 19 orchards for which we dispose of information), which may prevent some long distance transmissions. Then, for some of the trees sampled, we had both majority and minority genetic sequences (i.e. found in smaller quantities). For now, we only used the information of the majority sequences because the model only allowed accounting for a unique genetic sequence for one tree. We could modify a part of the model to take into account this information. To finish, it could be interesting to sample data over more than 3 years in order to really exploit the temporal information of the data.

ACKNOWLEDGMENTS

This work was supported by the CIRAD-UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.southgreen.fr). This work is supported by an INRA/ANSES scholarship to CP, and by INRA (BEcOSMASH project, funded by the SMaCH Metaprogram and the SAE2 Department) and Agropolis Fondation (E-SPACE project).

REFERENCES

Ashkenazy H., Penn O., Doron-Faigenboim A., Cohen O., Cannarozzi G., Zomer O. and Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40:W580–W584.

Cambra M., Capote N., Myrta A. and Llácer G. (2006). *Plum pox virus* and the estimated costs associated with sharka disease. EPPO Bull. 36(2):202-204.

Cunniffe N.J., Koskella B., E. Metcalf C.J., Parnell S., Gottwald T.R. and Gilligan, C.A. (2015). Thirteen challenges in modelling plant diseases. Epidemics 10:6–10.

Dallot S., Borron, S. Bertanpetit, E., Dupuy V., Jacquot E. and Thébaud, G. (2016). Exploiting viral genetic diversity to uncover sharka dispersal at regional and local scales. In: *Building bridges between disciplines for sustainable management of plant virus diseases. IPVE 2016. Programme and abstracts* (p. 23). Presented at 13. International plant virus epidemiology symposium, Avignon, FRA (2016-06-06 - 2016-06-10). 165 p. https://prodinra.inra.fr/record/361232.

Ferguson N.M., Keeling M.J., Edmunds W.J., Gani R., Grenfell B.T., Anderson R.M. and Leach, S. (2003). Planning for smallpox outbreaks. Nature 425(6959):681–685.

Gibson G.J. (1997). Markov Chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. J. R. Stat. Soc. Ser. C Appl. Stat. 46:215–233.

Jombart T., Cori A., Didelot X., Cauchemez S., Fraser C. and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol 10(1):e1003457.

Keeling,M.J., Woolhouse M.E.J., May R.M., Davies G. and Grenfell B.T. (2003). Modelling vaccination strategies against foot-and-mouth disease. Nature 421(6919):136–142.

Klinkowski M. (1970). Catastrophic plant diseases. Annu. Rev. Phytopathol. 8:37-60.

Lau M.S.Y., Marion G., Streftaris G., Gibson G. (2015). A systematic Bayesian integration of epidemiological and genetic data. PLOS Comput. Biol. 11(11):e1004633.

Mollentze N., Nel L.H., Townsend S., Roux K., Hampson K., Haydon D.T. and Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc. R. Soc. Lond. B Biol. Sci. 281(1782):20133251.

Morelli M.J., Thébaud G., Chadœuf J., King D.P., Haydon D.T. and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput Biol 8(11):e1002768.

Neri F.M., Cook A.R., Gibson G.J., Gottwald T.R. and Gilligan, C.A. (2014). Bayesian analysis for inference of an emerging epidemic: citrus canker in urban landscapes. PLOS Comput Biol 10(4):e1003587.

Ostfeld R.S., Glass G.E. and Keesing F. (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. Trends Ecol. Evol. 20(6):328–336.

Picard C., Dallot S., Brunker K., Berthier K., Roumagnac P., Soubeyrand S., Jacquot E. and Thébaud, G. (2017). Exploiting Genetic Information to Trace Plant Virus Dispersal in Landscapes. Annu. Rev. Phytopathol. Vol. 55 (in press).

Picard C., Soubeyrand S., Jacquot E. and Thébaud G. Analyzing the influence of landscape aggregation on disease spread to improve management strategies. In revision.

Picard C., Picheny V., Bonnot F., Soubeyrand S. and Thébaud G. In silico optimization of a strategy for landscape-wide plant disease management. In prep.

Pleydell D., Soubeyrand S., Dallot S., Labonne G., Chadœuf J., Jacquot E. and Thébaud, G. (2018). Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape. PLOS Comput. Biol. 14(4):e1006085.

Rimbaud L., Dallot S., Gottwald T., Decroocq V., Jacquot E., Soubeyrand S. and Thébaud G. (2015). Sharka epidemiology and worldwide management strategies: learning lessons to optimize disease control in perennial plants. Annu. Rev. Phytopathol. 53:357–378.

Rimbaud L., Dallot S., Bruchou C., Thoyer S., Jacquot E., Soubeyrand S. and Thébaud G. (2018). Heuristic optimisation of the management strategy of a plant epidemic using sequential sensitivity analyses. bioRxiv 315747. (doi: https://doi.org/10.1101/315747)

Soubeyrand S. (2016). Construction of semi-Markov genetic-space-time SEIR models and inference. J. Société Fr. Stat. 157(1):129–152.

Vurro M., Bonciani B. and Vannacci, G. (2010). Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences. Food Secur. 2(2):113–132.

Worby C.J., Lipsitch M. and Hanage W.P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLOS Comput. Biol. 10(3):e1003549.

Ypma R.J.F., Bataille A.M.A., Stegeman A., Koch G., Wallinga J. and Van Ballegooijen, W.M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc. R. Soc. Lond. B Biol. Sci. 279(1728):444–450.

Ypma R.J.F., Van Ballegooijen W.M. and Wallinga J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195(3):1055–1062.

S1 Text: Inference of transmission tree of real sharka data: materiel and method

We attempted to infer the transmissions of a real sharka epidemic. Here, we present the epidemiological and genetic data used, and we introduce the modifications made in the genetic-space-time model.

1. Epidemiological and genetic sharka data

We used data from 3 years of surveillance (2004, 2005 and 2006) of a very close set of 19 orchards located in southern France (Fig 1). Disease surveillance was based on visual inspections and PPV symptomatic trees were removed each year following their detection. The 4905 trees (among which 145 were found infected by PPV) planted in the 19 orchards were precisely geo-referenced. The plantation dates of each orchard were recorded as well as the dates of detection and removal of the infected trees. Symptomatic leaves were sampled on each detected PPV infected tree and we obtained the whole genome PPV consensus sequence for 114 infected trees. In addition, a preliminary phylogeny study evidenced that two different genetic clades of PPV were spreading in the study area (Dallot et al., 2016). The common ancestors of these two clades were reconstructed from the 145 genetic sequences sampled on infected trees thanks to FastML server (Ashkenazy et al., 2012). Here, we used the genetic-space-time model on data corresponding to only 6 patches (Fig 1).

Figure 1: Map of peach orchards with uninfected (grey) and infected trees detected in 2004 (yellow), 2005 (pink) and 2006 (blue). Only the orchards framed in black were used in the inference of transmission trees and pathogen dispersal.

2. Genetic-space-time SEIR model modifications

The genetic-space-time SEIR model was extended to handle the use of multiple ancestral sequences of the pathogen corresponding to different genetic clusters. Regarding the epidemiological and evolutionary events that are included in the model, if host *k* is infected by an exogenous source (this possibility depends on the basic risk α_0), then the ancestral sequence selected for the exogenous source is the ancestral sequence used for the genetic cluster to which the pathogen sequence collected from host *k* belongs.

In addition, we specified a new prior for parameters of the genetic-space-time SEIR model. We considered at least two introduction events from genetically differentiated PPV sources corresponding to the distinct clades in the reconstructed dated phylogeny of PPV. Based on BEAST inferences (Dallot et al., 2016), the time to the most recent common ancestor (tMRCA) of these two clades was set at 1985 (CI95% : 1981 – 1989, i.e., 19 years before the discovery of sharka disease in the area).

Résultats clés de l'Article 3

PRISE EN COMPTE DES HOTES NON INFECTES DANS LA RECONSTRUCTION DE CHAINES DE TRANSMISSION

- **Impact de la prise en compte des hôtes non infectés dans l'inférence des transmissions d'une épidémie**
  - Un modèle généticospatio-temporel déjà existant permettant d'inférer « qui a infecté qui » dans un paysage a été modifié pour prendre en compte les hôtes sensibles mais non infectés.
  - Grâce à la reconstruction d'épidémies simulées (et du processus évolutif concomitant), nous avons montré que la prise en compte des hôtes non infectés permettait d'inférer correctement 35% des transmissions (contre 20% sans).
  - Inclure les hôtes non infectés dans une analyse de données épidémiologiques et génétiques permet donc une meilleure compréhension de l'épidémiologie spatiale d'un agent pathogène et fournit des indications précieuses sur la dynamique de transmission. Une telle connaissance des transmissions est cruciale pour concevoir des politiques efficaces pour gérer les épidémies.

- **Reconstruction des chaines de transmission pour une épidémie de sharka**
  - Nous avons tenté de reconstruire les chaines de transmission pour une épidémie de sharka dans un ensemble de vergers proches. Nous disposions des coordonnées géographiques des arbres infectés et des séquences génétiques du virus correspondant.
  - Probablement à cause du manque de signal temporel, nous n'avons pas obtenu de résultats satisfaisants. Les paramètres du modèle représentant les épidémies de sharka n'ont donc pas été modifiés dans les études présentées dans la suite de cette thèse.