Phytopathology • 2017 • 107:1199-1208 • https://doi.org/10.1094/PHYTO-02-17-0070-FI

e-Xtra*

Testing Differences Between Pathogen Compositions with Small Samples and Sparse Data

Samuel Soubeyrand,† Vincent Garreta, Caroline Monteil, Frédéric Suffert, Henriette Goyeau, Julie Berder, Jacques Moinard, Elisabeth Fournier, Didier Tharreau, Cindy E. Morris, and Ivan Sache

First and second authors: BioSP, INRA, 84914, Avignon, France; second, fourth, fifth, and sixth authors: INRA, UMR1290 Bioger, AgroParisTech, Université Paris-Saclay 78850 Thiverval-Grignon, France; third and tenth authors: INRA, UR0407 Plant Pathology, 84143 Montfavet, France; seventh author: DRAAF Midi-Pyrénées, 31074 Toulouse Cedex, France; eighth author: INRA, UMR BGPI, 34398 Montpellier, France; ninth author: CIRAD, UMR BGPI, 34398 Montpellier, France; and eleventh author: AgroParisTech, UMR1290 Bioger, 78850 Thiverval-Grignon, France.

Accepted for publication 27 June 2017.

ABSTRACT

The structure of pathogen populations is an important driver of epidemics affecting crops and natural plant communities. Comparing the composition of two pathogen populations consisting of assemblages of genotypes or phenotypes is a crucial, recurrent question encountered in many studies in plant disease epidemiology. Determining whether there is a significant difference between two sets of proportions is also a generic question for numerous biological fields. When samples are small and data are sparse, it is not straightforward to provide an accurate answer to this simple question because routine statistical tests may not be exactly calibrated. To tackle this issue, we built a computationally intensive testing procedure, the generalized Monte Carlo plug-in test with calibration test, which is implemented in an R package available at https://doi.org/10.5281/zenodo.635791. A simulation study was carried out to assess the performance of the proposed methodology and to make a comparison with standard statistical tests. This study allows us to give advice on how to apply the proposed method, depending on the sample sizes. The proposed methodology was then applied to real datasets and the results of the analyses were discussed from an epidemiological perspective. The applications to real data sets deal with three topics in plant pathology: the reproduction of *Magnaporthe oryzae*, the spatial structure of *Pseudomonas syringae*, and the temporal recurrence of *Puccinia triticina*.

The genetic structure of pathogen populations is an important driver of epidemics affecting crops (García-Arenal et al. 2001; McDonald and Linde 2002) and natural plant communities (Burdon 1993; Gilbert 2002). The composition of a plant pathogen population can be defined as an array of genotypes or, when genotyping is not feasible, of the phenotypic expression of genotypes (Hull 2008). Determining whether the compositions of two populations of pathogens are different is a generic question that arises in many situations. The populations to be compared might have been sampled across various temporal scales within a single epidemic season (Villaréal and Lannou 2000) or across successive epidemic seasons (Tian et al. 2015); geographical scales from a single field to a country (Goyeau et al. 2012), a continent (Kolmer et al. 2012; Pule et al. 2013) or worldwide (Linde et al. 2002); and ecological niches such as different host plants (Leroy et al. 2014) or ecosystem components (Vinatzer et al. 2014).

The recent theoretical and methodological advances for analyzing microbial population genetics (Xu 2010) have been successfully applied to several plant pathogens (Milgroom 2015). Those methods rely on strong assumptions regarding the biology of the microorganisms, which are often violated for plant pathogens (Rozenfeld et al. 2007). First, sexual reproduction is a prerequisite, although several pathogens do not show any signature of such a reproduction and exhibit clonal population structure; the sexual system, moreover, is a matter of speculation in many plant pathogens (Tollenaere and Laine 2013). Second, the markers used for population comparison should be under selective neutrality, which is clearly not the case when the markers are linked with virulence (Gérard et al. 2006). Therefore, there is an important need for a statistical method for comparison of populations that does not require any a priori knowledge about pathogen biology.

Moreover, the sizes of populations of plant-pathogenic microorganisms are usually orders of magnitude greater than the size of the samples that can be taken and characterized from these populations. Thus, relatively small samples that result in sparse data about population structure are more the rule than the exception. In addition, sample size is particularly small in the following specific but frequent situations: when collection and identification of samples are costly and time consuming, and when a simple pilot study is carried out to determine the relevancy of performing a larger study and to design it. The problem of inherently small sample size is exacerbated by the genetic diversity of clonal microbial populations, where there are usually only a few dominant variants and multiple variants of exponentially decreasing abundance (Arnaud-Haond et al. 2007). Thus, for clonal microorganisms, not only are sample sizes small but data are also sparse for nondominant variants. Therefore, statistical tests adapted to small sample sizes and sparse data are needed.

For comparing vectors of proportions, the χ^2 test and the Fisher's exact test (Agresti 2007) are routinely used. However, both tests may not be exactly calibrated when sample sizes are small and data are sparse, even if the Monte Carlo versions of the tests (Hope 1968; Manly 1997) are used, because calculations are dependent on counts in the margins of the contingency table. A test is not calibrated if the actual risk of false rejection (i.e., type I error) is different from the significance level (e.g., 0.05) specified by the user (Sellke et al. 2001). Unconditional exact tests have been proposed to solve this issue for 2-by-2 contingency tables (i.e., with two groups

†Corresponding author: S. Soubeyrand; E-mail: samuel.soubeyrand@inra.fr

*The *e-Xtra* logo stands for "electronic extra" and indicates that five supplementary figures, three supplementary tables, and four supplementary text files are published online.

and two response types) (Lydersen et al. 2009; Mehrotra et al. 2003). These tests rely on a maximization with respect to the probabilities of all the response types. For a 2-by-2 contingency table, there are two response types with probabilities π and $1-\pi$ under the null hypothesis and, therefore, the maximization is carried out with respect to a single parameter (π). This maximization approach, adopted in unconditional exact tests, prevents their application for larger contingency tables. Indeed, if we consider a design with two groups and n response types (as in our case studies), then the maximization must be done with respect to $n-1$ probabilities, which is complicated as soon as n is large, given the limited information contained in a contingency table, especially when samples are small.

The main objective of this article is to describe and evaluate an alternative calibrated statistical procedure to test the similarity of compositions of two populations of a pathogen based on small samples and sparse data (typically, several dozen variants of the pathogen, including a large number of nondominant variants, and a few dozen isolates), without any a priori biological knowledge. This procedure has a wide spectrum of applications because, from a generic point of view, it aims at testing the equality of two unknown vectors of probabilities (p_1 and p_2) based on two multinomial draws performed with these probabilities.

Thus, after demonstrating that standard statistical tests are not calibrated in the case of small samples and sparse data, we propose a new test based on a numerical calibration. Then, simulation studies comparing the performances of the standard tests and the proposed new one are provided. The comparisons are based on type I error and the power of the tests when the sample sizes and the frequencies of variants in the samples vary. Following the conclusions of the simulation study, the new test was applied to real datasets raising issues of theoretical and practical relevance in plant disease epidemiology: (i) the reproduction regime of a rice fungal pathogen, *Magnaporthe oryzae*, in Madagascar and China; (ii) the population diversity of a bacterial pathogen, *Pseudomonas syringae*, from alpine areas to crops in southeastern France; and (iii) the temporal recurrence of a wheat fungal pathogen, *Puccinia triticina*, in southwestern France.

The data sets and the computer code for applying the method are provided in the generalized Monte Carlo plug-in test with calibration (GMCPIC) R package, available at https://doi.org/10.5281/zenodo.635791.

MATERIALS AND METHODS

Why routine tests might not necessarily be calibrated? Consider two vectors of counts, say N_1 and N_2, independently drawn under multinomial distributions with unknown vectors of probabilities p_1 and p_2. With the aim of testing the equality $p_1 = p_2$ of the vectors of probabilities using N_1 and N_2, the χ^2 test and the Fisher's exact test are routinely applied. However, the sparseness of N_1 and N_2 hampers the use of these tests, even when the P value is computed using Monte Carlo simulations. To understand this statement, one can inspect the formula of the χ^2 statistic Q used in the χ^2 test:

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{K} \frac{\left(N_{ij} - n_i \hat{q}_j\right)^2}{n_i \hat{q}_j} \quad (1)$$

where K is the number of categories, N_{ij} is the jth component of N_i, $j \in \{1,\ldots,K\}$ and $i \in \{1,2\}$, n_i is the size of sample i (i.e., $n_i = \sum_j N_{ij}$), and $\hat{q}_j$ is the proportion of items from category j in both samples. Note that $\hat{q}_j$ is equal to the following weighted means:

$$\hat{q}_j = \frac{n_1}{n_1 + n_2}\hat{p}_{1j} + \frac{n_2}{n_1 + n_2}\hat{p}_{2j} \quad (2)$$

where $\hat{p}_{1j}$ and $\hat{p}_{2j}$ are the proportions of items from category j in samples 1 and 2, respectively.

When the contingency table (i.e., the matrix with columns N_1 and N_2) is sparse, under the null hypothesis, the estimates $n_i\hat{q}_j$ are unbiased but relatively strongly varying estimates of the expectations $E(N_{ij})$ of N_{ij}—that is, the ratios $n_i\hat{q}_j / E(N_{ij})$ vary strongly. Therefore, when the classical χ^2 test is applied, the normal approximation of the distribution of $n_i\hat{q}_j$ is crude and so is the χ^2 approximation of the distribution of the statistic Q. Similarly, when the Monte Carlo version of the χ^2 test is applied, the simulated counts replacing the observed counts N_{ij} are obtained under the probabilities $\hat{q}_j$ that are significantly different from the true probabilities under which the observed counts were generated. Consequently, the Monte Carlo approximation of the distribution of the statistic Q is crude. Such a crude approximation of the distribution of Q leads to a calculated P value that does not exactly give the probability $P_{H_0}(Q > Q_{obs})$ of observing Q at least as extreme as the observed value Q_{obs} of Q under the null hypothesis H_0. If the difference between the calculated P value and its theoretical counterpart $P_{H_0}(Q > Q_{obs})$ is not negligible, then the test is uncalibrated. A test is calibrated if the actual risk of false rejection (i.e., type I error) is equal to the significance level α specified by the user (Sellke et al. 2001), and the significance level α is a threshold under which the P value of the test is considered statistically significantly low and leads to the rejection of the null hypothesis.

The same argument can be used for the Fisher's exact test, where the counts observed in the margins of the contingency table are used to specify the distribution of the counts observed inside the contingency table, and for the Monte Carlo plug-in test detailed in Supplementary Text S1.

For 2-by-2 contingency tables, unconditional exact tests have been proposed to solve the issue mentioned above (Lydersen et al. 2009; Mehrotra et al. 2003) and can be easily applied, for example, with the Exact and Barnard R packages. However, for larger tables such as those considered in this article, no routine test exists.

GMCPIC test. As seen above, with sparse data, the inadequacy of the χ^2 test, the Fisher's exact test, and the Monte-Carlo plug-in test is due to a relatively strongly varying estimate $\hat{p}$ of the unknown vector of probabilities p of the multinomial distributions appearing in the null hypothesis. Under the null hypothesis, $N_1 \sim$ multinomial (n_1,p_1), $N_2 \sim$ multinomial(n_2,p_2), and $p = p_1 = p_2$. For example, in the χ^2 test, one uses the maximum-likelihood estimate of p based on the two samples:

$$\hat{p} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2 \quad (3)$$

The relatively strong variations of $\hat{p}$ with small sample sizes lead to uncalibrated tests; that is to say, tests whose significance levels are not satisfied in practice.

Here, we propose a Monte Carlo test based on a statistic $S(N_1, N_2, w)$ depending on a generalized version $\hat{p}(w)$ of the estimates $\hat{p}$ of p. The generalized estimate $\hat{p}(w)$ is a weighted mean of $\hat{p}_1$ and $\hat{p}_2$ that depends on a weight w belonging to the interval [0,1]:

$$\hat{p}(w) = w\hat{p}_1 + (1-w)\hat{p}_2 \quad (4)$$

and the weight w is selected such that the resulting test is calibrated at a fixed significance level α. Without loss of generality, the statistic $S(N_1,N_2,w)$ is expected to be large if the null hypothesis is true and small otherwise.

Suppose that the weight w has been selected, the generalized Monte Carlo plug-in test based on N_1 and N_2 is implemented as follows:

- independently draw $2B$ samples $N_1^{(b)}$ and $N_2^{(b)}$ ($b \in \{1,\ldots,B\}$, B large) under the multinomial distributions with sizes n_1 and n_2, respectively, and with vector of probabilities $\hat{p}(w)$;

- compute the P value $pval(N_1,N_2,w)$ of the test as the proportion of statistics $S(N_1^{(b)},N_2^{(b)},w)$ less than or equal to $S(N_1,N_2,w)$:

$$pval(N_1,N_2,w)=\frac{1}{B}\sum_{b=1}^{B} I\left\{S\left(N_1^{(b)},N_2^{(b)},w\right)\le S(N_1,N_2,w)\right\} \quad (5)$$

where $I\{E\}=1$ if event E occurs and $I\{E\}=0$ otherwise.

The selection of w (in other words, the calibration of the test) is carried out as follows:

- independently draw $2M$ samples $\tilde{N}_1^{(m)}$ and $\tilde{N}_2^{(m)}$ ($m\in\{1,\ldots,M\}$, M large) under the multinomial distributions with sizes n_1 and n_2, respectively, and with vector of probabilities $\hat{p}\left(\frac{n_1}{n_1+n_2}\right)$ corresponding, under the null hypothesis, to the maximum-likelihood estimate of p based on the two samples;
- minimize the following calibration criterion with respect to w depending on the fixed significance level α:

$$\left|\alpha-\frac{1}{M}\sum_{m=1}^{M} I\left\{pval\left(\tilde{N}_1^{(m)},\tilde{N}_2^{(m)},w\right)\le\alpha\right\}\right| \quad (6)$$

and let $\tilde{w}$ denote the minimizer of this criterion.

The GMCPIC test can be applied with various statistics, especially the extension of the negative χ^2 statistic:

$$S(N_1,N_2,\tilde{w})=-\sum_{i=1}^{2}\sum_{j=1}^{K}\frac{\left(N_{ij}-n_i\hat{p}_j(\tilde{w})\right)^2}{n_i\hat{p}_j(\tilde{w})} \quad (7)$$

and the extension of the statistic used in the plug-in test without calibration:

$$S(N_1,N_2,\tilde{w})=m(N_2;n_2,\hat{p}(\tilde{w})) \quad (8)$$

The GMCPIC test can be viewed as an intermediate between conditional and unconditional tests: (i) the test is still "conditional" because the vector of probabilities of response types under the null hypothesis is estimated by $\hat{p}(\tilde{w})$, defined as a weighted mean of the observed probability vectors $\hat{p}_1$ and $\hat{p}_2$, but (ii) the estimate $\hat{p}(\tilde{w})$ is obtained via a numerical maximization, such as in unconditional tests. However, in contrast to unconditional exact tests, the maximization is made with respect to a single parameter, whatever the number of response types. This point makes the GMCPIC test applicable to high-dimension vectors of counts N_1 and N_2 but is also the reason why the GMCPIC test is only approximately calibrated. To improve the calibration, the estimate $\hat{p}$ of the vector of probabilities under the null hypothesis should be searched for in a larger space. However, there is a trade-off between calibration and computation time. This topic is evoked again in the Discussion.

Remark: in the procedure described above, w is selected such that the test is calibrated or, in other words, such that the calculated P value is, under the null hypothesis, lower than the significance level α with a probability α (this is the meaning of minimizing the criterion given by equation 6). Minimizing this criterion is not equivalent, in general, to maximizing the likelihood for the model "$N_1\sim$ multinomial (n_1,p), $N_2\sim$ multinomial (n_2,p), $N_1\perp N_2$", which leads to the maximum-likelihood estimate

$$\hat{p}=\frac{n_1}{n_1+n_2}\hat{p}_1+\frac{n_2}{n_1+n_2}\hat{p}_2$$

of p. Thus, the minimizer $\tilde{w}$ of equation 6 is not, in general, equal to $n_1/(n_1+n_2)$, as we will see in the simulation studies presented in this article.

Simulation design for assessing type I errors. We numerically assessed type I errors of the tests mentioned above by applying them to several types of data sets generated under the null hypothesis (equality of p_1 and p_2). This numerical study was carried out with varying sample sizes ($n_1=n_2=$ 10, 100, or 1,000) and varying numbers of categories (3 or 33; this is the dimension of N_1 and N_2). For vectors with three categories, we used either homogeneous probabilities (1/3, 1/3, and 1/3) or heterogeneous probabilities (0.80, 0.19, and 0.01). For vectors with 33 categories, we used heterogeneous probabilities (0.70, 0.10, 0.10, 0.10/30,…, 0.10/30). In total, 1,000 data sets were generated in each case. The performances of the tests were assessed by computing the type I error (i.e., the incorrect rejection rate of the true null hypothesis) at the tolerance threshold 0.05.

The simulation series with 33 categories, heterogeneous probabilities, and small sample sizes are supposed to mimic typical data sets that are handled when one compares the compositions of two populations of pathogens. However, the GMCPIC test performance has to be assessed in other settings to evaluate whether it is a relevant alternative to standard tests when sample sizes are small, whatever the context in which the test is applied. The two simulation series with three categories were run for this purpose. To complete these series, we provided a more generic simulation study where vectors of probabilities are randomly generated with varying means and variances (see below).

Simulation design for assessing the power of the tests. The powers of the tests were numerically assessed by applying the tests to data sets generated under 12 different alternative hypotheses (inequality of p_1 and p_2) and by computing, for each alternative, the rate of rejection of the null hypothesis (the higher the rejection rate, the larger the test power). In this study, we used vectors of counts with 33 categories and with varying sample sizes ($n_1=n_2=$ 10, 100, or 1,000). In each simulation, N_1 was drawn with vector of probabilities $p_1=$ (0.70, 0.10, 0.10, 0.10/30,…,0.10/30) and N_2 was drawn with p_2 equal to one of the four following vectors of probabilities:

Modification type 1: (0.70 + δ, 0.10, 0.10, 0.10/30,…,0.10/30)/(1 + δ),
Modification type 2: (0.70, 0.10 + δ, 0.10, 0.10/30,…,0.10/30)/(1 + δ),
Modification type 3: (0.70, 0.10, 0.10, 0.10/30 + δ,…,0.10/30)/(1 + δ), and
Modification type 4: (0.70 − δ, 0.10, 0.10, 0.10/30 + δ,…,0.10/30),

where the amplitude δ of the difference is equal to either 0.2, 0.4, or 0.6 (the higher δ, the larger the difference between the two vectors of probabilities) (Supplementary Fig. S1). The first three modifications correspond to an increase of one of the categories (either the main category, a significant category, or a rare category) and, as compensation, a decrease of all other categories. In the fourth modification, there is an increase of one of the rare categories affecting only the main category that becomes less dominant. In all, 1,000 data sets were generated for each sample size value and each of the 12 alternative hypotheses (4 forms of p_2 for 3 values of the amplitude δ).

With the aim of assessing the performances of the tests for more diverse alternative hypotheses, Supplementary Table S1 provides a complementary assessment of the powers of the tests when vectors of probabilities are randomly generated with varying means and variances.

Applications to real datasets. In the following applications, vectors of counts N_1 or N_2 are compositions of pathogen populations (pathogen compositions [PC]). A PC is defined as a vector of frequencies of different variants of the pathogen found in a sample of isolates. Below, a variant designates either a multilocus genotype (*M. oryzae*), a virulence phenotype (*P. triticina*), a haplotype, a clade, or a phylogroup (*Pseudomonas syringae*). For highly diverse pathogens, the number of different variants that are considered may be large, and zeros in vector describing the PC may be frequent if the sample size is moderate.

The following paragraphs present the data sets that are analyzed in this article. These data sets are provided in the GMCPIC R package. Additional details are provided in Supplementary Text S2.

M. oryzae. Two data sets were collected in Madagascar and China for studying the reproduction regime of rice fungal pathogen

M. oryzae. In Madagascar, aerial organs of rice plants infected by *M. oryzae* were sampled on experimental upland rice plots from a single variety (Saleh et al. 2014) in February and April 2005. In Yunnan (China), infected panicles were collected in the same place in August 2008 and September 2009. These samples correspond to populations CH1-2008 and CH1-2009, respectively, described by Saleh et al. (2012). Sample locations are shown in Supplementary Figure S2. Field samples were purified by subculturing from single spores (the technique called "monosporing") and the resulting strains were genotyped according to 13 microsatellite markers (Saleh et al. 2012). For each strain, the combination of data from all the markers defined the multilocus genotypes (MLG) and, for each population, the number of strains per MLG was counted. Therefore, the PC (summarized in Table 1; detailed in Figure 1) consist of frequencies of MLG in each population.

This application was selected as a means to validate our procedure: we expect that the similarity of PC will be rejected in China, where partial sexuality is known to occur and the resulting recombination will lead to a change in the frequencies of the MLG. Alternatively, we expect that the hypothesis will not be rejected in Madagascar, where reproduction is known to be strictly clonal and where no bottleneck has arisen between the two sampling dates. Indeed, the Madagascar populations are clonal, in that the fungus multiplies by asexual spores only (Saleh et al. 2012, 2014). Under clonality, the frequencies of MLG are expected to be stable in time. In contrast, the Chinese population is reproducing sexually. Evidence was provided by genotyping a population sampled for two consecutive years in the same place, supplemented with biological data and simulations (Saleh et al. 2012). Recombination occurring during sexual reproduction leads to reassortment of allelic associations, thus creating new MLG. Thus, frequencies of MLG vary between two samples separated by at least one event of sexual reproduction.

TABLE 1. Number of *Magnaporthe oryzae* isolates and number of different variants in each sample[a]

Statistic	Data set	Samples		
		First	Second	Pooled
Number of isolates	China	24	83	107
	Madagascar	17	40	57
Number of different variants	China	21	72	92
	Madagascar	10	12	18

[a] For *M. oryzae* sampled in China, the first sample was collected in August 2008 and the second in September 2009. For *M. oryzae* sampled in Madagascar, the first sample was collected in February 2005 and the second in April 2005.

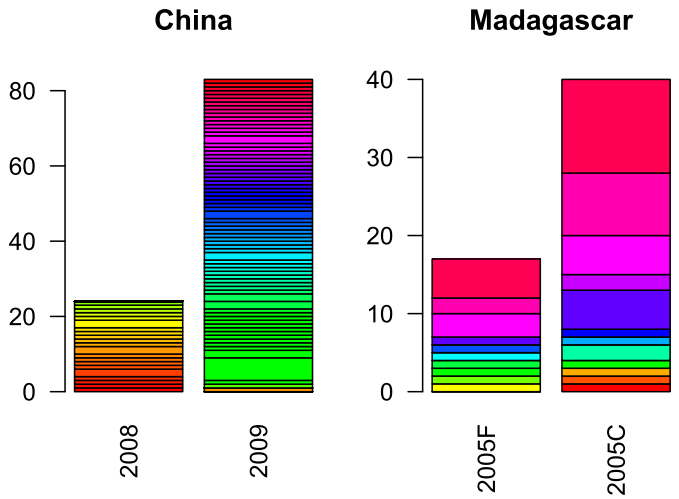

Fig. 1. Compositions of populations of *Magnaporthe oryzae* corresponding to samples collected in China (first sample collected in August 2008; second sample collected in September 2009) and Madagascar (first sample collected in February 2005; second sample collected in April 2005). Each colored layer corresponds to a given variant; the height of each layer is proportional to the number of isolates from the corresponding variant.

P. syringae. Genotypic data of *P. syringae* populations were collected from precipitation in two different but connected environments (namely, the southern French Alps and the agricultural lands irrigated downstream by the Durance River) to investigate the spatial diversity of *P. syringae* populations (Monteil et al. 2014a). Both environments may be connected because members of *P. syringae* strains are able to disseminate through air and water fluxes (Monteil et al. 2014a,b). The PC taken from the Alps and the PC taken from the crops were considered at three different resolutions: haplotypes, clades, and phylogroups (Berge et al. 2014). Our testing procedure was applied to PC observed in both areas to determine whether the diversity of *P. syringae* considered at various resolutions significantly differs between markedly contrasted ecosystems: those dominated by agriculture in the Lower Durance River (LDR) basin, downstream in the plains joining the Rhône River, and those dominated by forests and mown meadows in the mountains of the Upper Durance River (UDR) basin in the French Alps. The null hypothesis was that PC in the LDR and UDR were the same due to mixing through air masses and water flow.

P. syringae samples were collected over 4 years in rainwater in 10 different sites of the LDR and UDR basins. Haplotypic strains were purified as described by Morris et al. (2008) and clustered in phylogroups and clades, using the sequence of the *cts* gene, as described by Morris et al. (2010). Two strains with a dissimilarity rate lower that 4.9% were assigned to the same phylogroup and were assigned to the same clade if their dissimilarity rate was lower than 2.0% (Berge et al. 2014). Therefore, the PC (summarized in Table 2; detailed in Figure 2) consisted of frequencies of haplotypes, clades, and phylogroups in the UDR and LDR basins.

Puccinia triticina. The temporal recurrence of *Puccinia triticina* was investigated by collecting leaves infected with *P. triticina* for seven consecutive years (2007 to 2013) from wheat fields located in southwestern France. Every year, two sentinel plots (each of them grown with the same variety) were sampled three to four times on dates depending on disease development (Table 3). At each sampling date, a maximum of 30 diseased leaves were collected from each plot. Samples were also collected from wheat volunteers (i.e., self-set wheat plants established as weeds from the previous growing season) once a year during the intercrop season, shortly before sowing the next wheat crop (Table 3), in plots previously grown with wheat in a radius of 10 km around the two aforementioned sentinel plots. A maximum of 10 infected volunteer leaves was collected from each surveyed plot, yielding one observed phenotypic composition per sampling date.

Our testing procedure was applied to compositions observed at successive sampling dates to study (i) temporal disruptions and continuations in the genetic structure of the local pathogen population and, more specifically, (ii) the role of wheat volunteers

TABLE 2. Number of *Pseudomonas syringae* isolates and number of different variants in each sample[a]

Statistic	Resolution	Samples		
		First	Second	Pooled
Number of isolates	All resolutions	100	110	210
Number of different variants	Phylogroups	5	8	10
	Clade	9	14	17
	Haplotype	19	44	57

[a] For *P. syringae*, the first sample was collected in the Upper Durance River basin (alpine samples) and the second in the Lower Durance River basin (agriculture samples); variants are defined with respect to three different resolutions: phylogroups, clades, and haplotypes.

in the yearly recurrence of disease in wheat crops. The null hypothesis was that oversummering of the pathogen on volunteers led to local perpetuation of disease over the whole period of study; accordingly, a single, multiyear epidemic would have occurred rather than successive yearly epidemics reinitiated every year.

Field samples were purified and the virulence of strains was determined according to standard techniques (Goyeau et al. 2006). These virulence phenotypes (pathotypes) were determined by inoculating a susceptible control cultivar and a set of 18 wheat cultivars differing in the factors that determine their resistance to

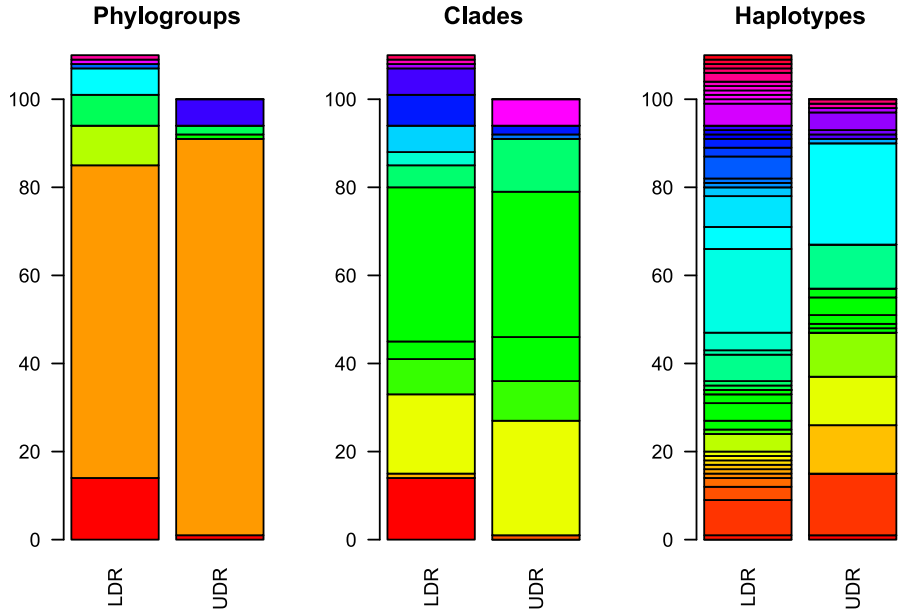

Fig. 2. Compositions of populations of *Pseudomonas syringae* corresponding to samples collected in Upper Durance River (UDR) basin and in Lower Durance River (LDR) basin considered at three different resolutions (phylogroups, clades, and haplotypes). Each colored layer corresponds to a given variant; the height of each layer is proportional to the number of isolates from the corresponding variant.

TABLE 3. Number of *Puccinia triticina* isolates sampled on Galibier and Kalango wheat at sampling dates during the cropping season (A, B, C, and D) and on wheat volunteers in the intercrop season (V) in years 2007 to 2013, and numbers of different variants

			Isolates[a]			Pathotypes[b]		
Year	Sample	Sampling date	Galibier	Kalango	Total	Galibier	Kalango	Both
2007	A	22 February	24	19	43	9	6	12
	B	4 April	11	17	28	6	6	10
	C	14 May	24	22	46	9	10	14
	V[c]	…	…	…	…	…	…	…
2008	A	1 April	0[d]	28	28	0	8	8
	B	5 May	24	18	42	4	8	10
	C	26 May	28	30	58	6	8	13
	V	28 October	10	10	20	4	3	6
2009	A	5 May	17	13	30	8	5	11
	B	25 May	30	28	58	1	8	9
	C	10 June	27	29	56	5	7	11
	V	27 October	0[d]	0[d]	0[d]	0	0	0
2010	A	4 May	11	10	21	4	4	6
	B	26 May	29	27	56	9	11	16
	C	15 June	30	30	60	9	12	18
	V	26 October	5	10	15	2	6	7
2011	A	21 April	23	28	51	10	6	12
	B	27 May	29	29	58	11	8	13
	C[e]	…	…	…	…	…	…	…
	V	26 October	10	10	20	7	3	8
	A	29 December	21	27	48	6	3	7
2012	B	29 March	0[d]	8	8	0	2	2
	C	25 April	10	8	18	4	4	6
	D	24 May	30	26	56	13	5	14
	V	6 November	11	13	24	8	5	11
2013	A	24-Mar	30	30	60	14	11	19
	B	17 May	30	30	60	12	7	14
	C	13 June	30	30	60	9	8	12
Pooled	…	…	494	530	1,024	50	37	64

[a] Number of isolates for which the pathotype was determined.
[b] Number of different pathotypes.
[c] No sampling on volunteers.
[d] No infected leaf was found.
[e] No third sampling on Galibier and Kalango in 2011.

P. triticina. Infection types on the differentials were evaluated 10 days after inoculation to establish the virulence phenotype of each strain. Therefore, the PC (summarized in Table 3; detailed in Figure 3) consisted of frequencies of virulence phenotypes at each sampling date.

RESULTS

Simulation-based study: analysis of type I error and power. Assessments of type I errors (i.e., the incorrect rejection of the true null hypothesis) in different settings are shown in Table 4. The χ^2 test, the Fisher's exact tests, and their Monte Carlo versions have incorrect type I errors when frequencies of variants are heterogeneous and sample sizes are small: they tend to under-reject the null hypothesis (i.e., they are conservative). It has to be noted that the Monte Carlo χ^2 test and the two Fisher's tests lead to very close type I errors. The Monte Carlo plug-in test consistently over-rejects the null hypothesis in any setting and is definitely an incorrect test. The GMCPIC test based on equation 7 shows an incorrect type I error for the large PC at small and moderate sample sizes (trend to under-rejection), whereas the GMCPIC test based on equation 8, which is a calibrated version of the Monte Carlo plug-in test, has correct type I errors in every setting. This difference in the two GMCPIC tests shows the importance of the choice of the statistics to be calibrated.

Assessments of powers (i.e., the correct rejection of the false null hypothesis) for PC with 33 variants, with varying type of difference between the two PC, and with varying amplitude δ of the difference, are compared (Fig. 4). First, it has to be noted that the Monte Carlo χ^2 test and both Fisher's tests have similar powers in all settings (however, the Fisher's exact test was not run for samples with size 1,000 because of excessive computation time). Second, the GMCPIC test based on equation 8 (Fig. 4, turquoise), which provided the most satisfactory results with respect to type I errors, has, for small sample sizes, a slightly better performance in rejecting the false null hypothesis than the three previously mentioned tests, which are not calibrated at small sample sizes. For larger sample sizes, the power of the GMCPIC test based on equation 8 can be lower than the power of the Monte Carlo χ^2 test and the two Fisher's tests, especially when the modification affects the dominant or a significant variant.

These results concerning the type I error and the power are corroborated by the results of the complementary simulation study, where the vectors of probabilities p_1 and p_2 are randomly generated with varying means and variances.

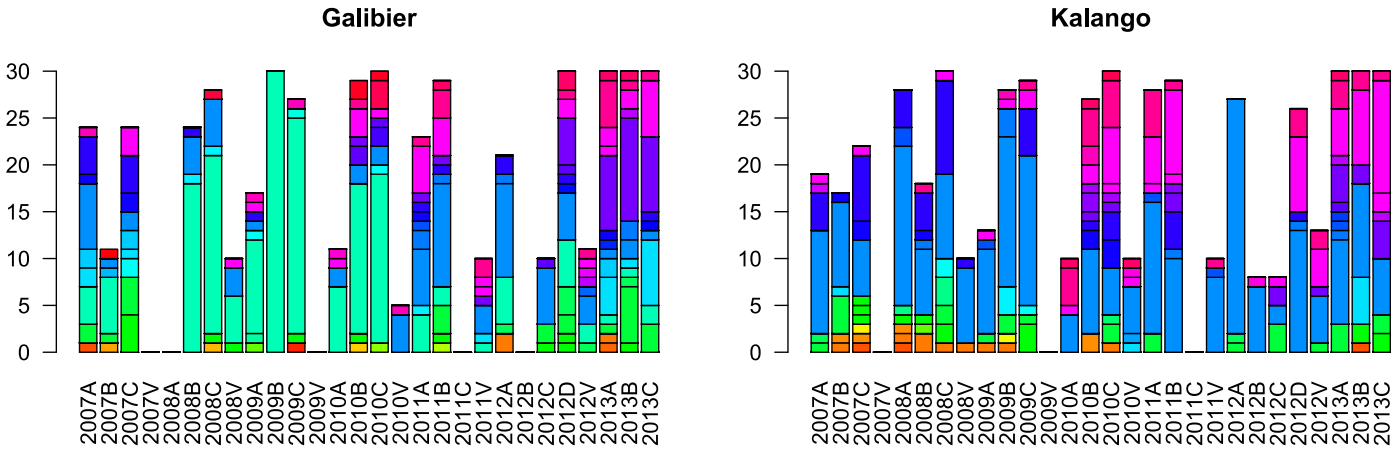

Fig. 3. Compositions of populations of *Puccinia triticina* sampled across time in Galibier and Kalango crops in southwestern France from 2007 to 2013. Letters A, B, C, D, and V refer to different sampling dates in each year of the study period. Each colored layer corresponds to a given variant; the height of each layer is proportional to the number of isolates from the corresponding variant.

TABLE 4. Type I errors of the χ^2 test and its Monte Carlo (MC) version (with $B = 10^4$ simulations), the Fisher's exact test and its MC version (with $B = 10^4$ simulations), the MC plug-in test (with $B = 10^4$ simulations), and the generalized Monte Carlo plug-in test with calibration (GMCPIC) test using the statistics of equation (7) or the statistics of equation 8 and using $B = 10^4$ and $M = 10^3$ simulations[a]

	Tests					GMCPIC test with statistics of	
PC, size[b]	χ^2	MC χ^2	Fisher's	MC Fisher's	MC plug-in	Equation 7	Equation 8
Hom, 3[c]							
10	0.039 (0.128)	0.036 (0.050)	0.037 (0.069)	0.038 (0.095)	0.324 (<0.0001)	0.044 (0.425)	0.042 (0.276)
100	0.052 (0.828)	0.052 (0.828)	0.051 (0.942)	0.050 (0.942)	0.246 (<0.0001)	0.055 (0.514)	0.050 (1.000)
1,000	0.058 (0.277)	0.056 (0.425)	0.055 (0.514)	0.056 (0.425)	0.232 (<0.0001)	0.053 (0.717)	0.056 (0.425)
Het, 3[d]							
10	0.008 (<0.0001)	0.008 (<0.0001)	0.008 (<0.0001)	0.008 (<0.0001)	0.248 (<0.0001)	0.046 (0.612)	0.045 (0.514)
100	0.044 (0.425)	0.054 (0.612)	0.053 (0.717)	0.055 (0.514)	0.357 (<0.0001)	0.053 (0.717)	0.057 (0.346)
1,000	0.048 (0.828)	0.049 (0.942)	0.048 (0.828)	0.046 (0.612)	0.219 (<0.0001)	0.042 (0.277)	0.046 (0.612)
Het, 33[e]							
10	0.005 (<0.0001)	0.024 (0.0002)	0.022 (<0.0001)	0.022 (<0.0001)	0.803 (<0.0001)	0.023 (0.0001)	0.056 (0.425)
100	0.006 (<0.0001)	0.051 (0.942)	0.052 (0.828)	0.054 (0.612)	1.000 (0.0000)	0.030 (0.005)	0.057 (0.346)
1,000	0.033 (0.017)	0.046 (0.612)	NA	0.051 (0.942)	0.973 (<0.0001)	0.053 (0.717)	0.042 (0.277)

[a] Type I error was computed for each type of pathogen composition and each sample size as the proportion of rejections over 1,000 repetitions. *P* values of the test of equality of the type I errors to the value 0.05 are shown in parentheses. For 33 variants and samples of size 1,000, the Fisher's exact test was not run because of excessive computation time. NA = not available.
[b] Pathogen composition (PC) and sample size.
[c] Three variants with homogeneous probabilities.
[d] Three variants with heterogeneous probabilities.
[e] Thirty-three variants with heterogeneous probabilities.

Thus, in the applications, we apply the GMCPIC test based on equation 8 (with $B = 10^4$ and $M = 10^3$), the Monte Carlo χ^2 test, and both Fisher's tests. For *M. oryzae* and *Pseudomonas syringae*, sample sizes are moderate (Tables 1 and 2) and we expect that the four tests will provide similar results. For *Puccinia triticina* data, sample sizes that range from 5 to 30 are low (Table 3) and we expect eventual differences in test results.

Reproduction of *M. oryzae*. For *M. oryzae* data, the tests were applied to a pair of PC sampled in China in August 2008 and September 2009 and to a pair of PC sampled in Madagascar in February 2005 and April 2005.

The similarity of PC separated in time is rejected for the Chinese population studied, where sexual reproduction was demonstrated to have occurred between the two sampling dates. In contrast, the similarity of PC separated in time and on different organs is not rejected for Madagascar data, where reproduction is known to be strictly clonal and where no bottleneck is expected between the two sampling dates (Table 5). The GMCPIC test based on equation 8 and the three other tests provide the same results for such sample sizes and such PC structures.

Spatial structure of *Pseudomonas syringae* populations. For *Pseudomonas syringae* data, the tests were applied to the samples collected in the UDR and LDR basins. Three resolutions of the samples were considered: variants were either phylogroups, clades, or haplotypes.

The four testing procedures reject the similarity of PC sampled in UDR and LDR basins at the three resolutions under consideration (Table 5). Thus, precipitation in the Durance River basin deposits populations whose diversity is different according to the area (agricultural or alpine).

Temporal recurrence of *Puccinia triticina*. For *Puccinia triticina* data, the tests were applied to each pair of consecutive samples collected only in fields sown with 'Galibier' wheat and to each pair of consecutive samples collected only in fields sown with 'Kalango' wheat. The tests were also applied to each pair of consecutive samples by merging data collected in fields sown with Galibier and Kalango and by discarding pathotypes that are not virulent for both Galibier and Kalango.

The periods at which the temporal continuation in the genetic structure of the local *P. triticina* population (i.e., the null hypothesis)

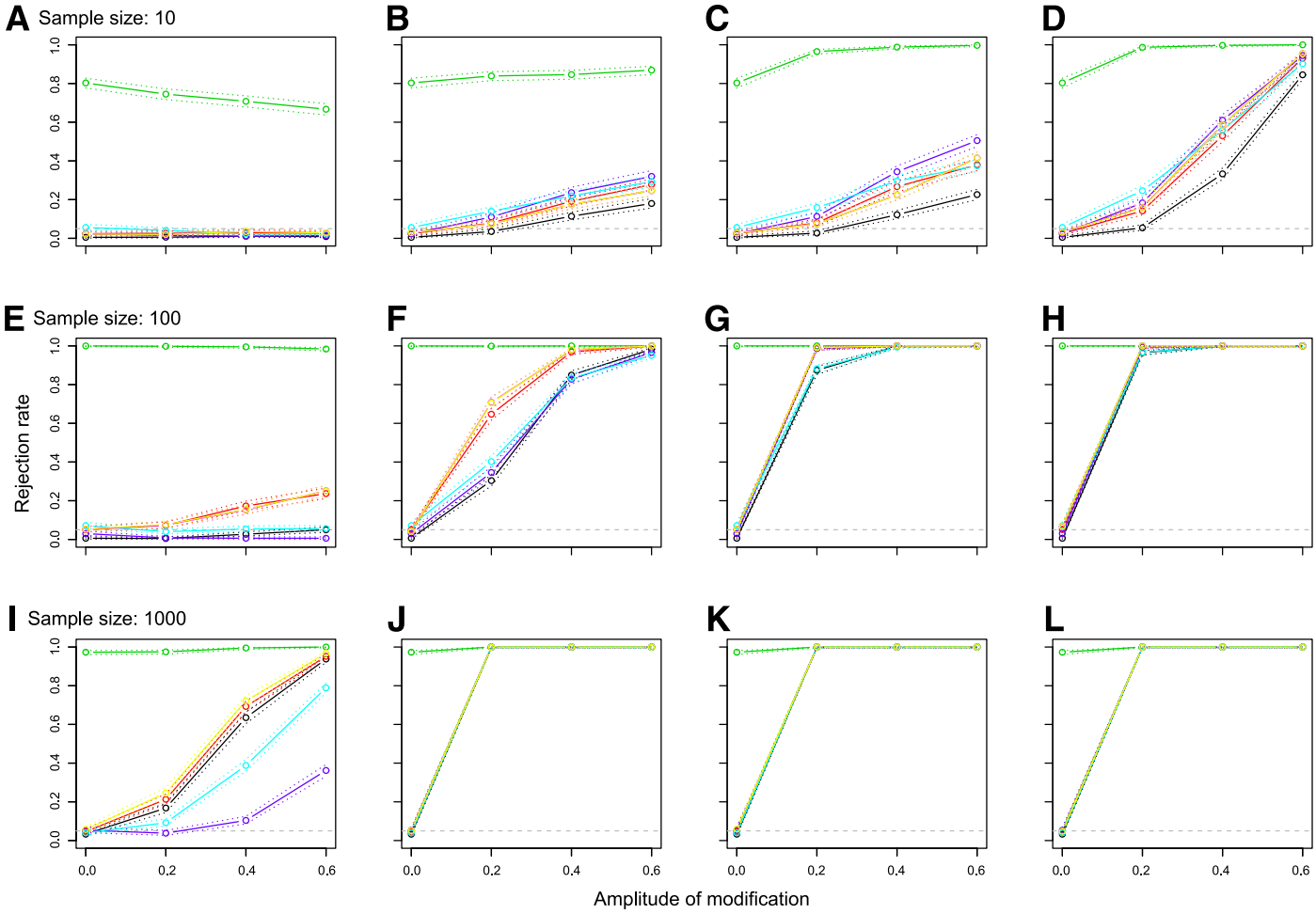

Fig. 4. Variation in the powers of the tests with respect to the difference between the vectors of probabilities p_1 and p_2 (the difference between p_1 and p_2 depends on the modification type and the amplitude δ of the modification). Each panel corresponds to a specific sample size and a specific modification type. In each panel, the colored curves give the variation in the powers of the following tests: χ^2 test (black) and its Monte Carlo version with $B = 10^4$ simulations (red), the Fisher's exact test (pink) and its Monte Carlo version with $B = 10^4$ simulations (yellow), the Monte Carlo plug-in test with $B = 10^4$ simulations (green), and the generalized Monte Carlo plug-in test with calibration test using the statistics of equation (7) (blue) and the statistics of equation 8 (turquoise) with $B = 10^4$ and $M = 10^3$ simulations. The powers were assessed over 1,000 repetitions for each modification type and each sample size. In all panels, p_1 = (0.70, 0.10, 0.10, 0.10/30,…,0.10/30). **A, E,** and **I** (modification type 1): p_2 = (0.70 + δ, 0.10, 0.10, 0.10/30,…,0.10/30)/(1 + δ); **B, F,** and **J** (modification type 2): p_2 = (0.70, 0.10 + δ, 0.10, 0.10/30,…,0.10/30)/(1 + δ); **C, G,** and **K** (modification type 3): p_2 = (0.70, 0.10, 0.10, 0.10/30 + δ,…,0.10/30)/(1 + δ); and **D, H,** and **L** (modification type 4): p_2 = (0.70 − δ, 0.10, 0.10, 0.10/30 + δ,…,0.10/30), where the amplitude δ of difference takes four different values (see x-axis). When δ = 0, the pathogen compositions are drawn under the same vectors of probabilities, and the corresponding rejection rate is a type I error. In each panel, the horizontal dashed gray line indicates the value 0.05 of the significance level, and the dotted envelopes give 95% confidence envelopes of the powers (pointwise assessments based on the binomial variation around the estimated powers). For samples of size 1,000 (bottom panels), the Fisher's exact test was not run because of excessive computation time.

is rejected by the GMCPIC test based on equation 8 are shown in Figure 5; that is to say, when there are disruptions in the PC (Supplementary Table S2 provides the corresponding P values). The total continuation of the epidemic with constant composition over the study period (2007 to 2013) is rejected for both Galibier and Kalango. Indeed, for Galibier and Kalango, 30 and 25% of the tests, respectively, reject the null hypothesis at the 5% significance level. Disruptions are mostly simultaneous in Galibier and Kalango crops. In addition, the disruptions can occur during the intercrop season (when *P. triticina* is thought to survive on volunteer wheat) but also during the crop season (when *P. triticina* is thought to reinfect the wheat crops).

Results obtained with the GMCPIC test based on equation 8, the Monte Carlo χ^2 test, and both Fisher's tests are compared in Supplementary Table S3. The GMCPIC test differs from the three other tests for nearly 10 comparisons of PC over 68 comparisons made in total. This relatively large difference between the tests is due, in this application, to the small sizes of the samples. Based on the simulation study presented above, the GMCPIC test is expected to provide, for this application, the more accurate results.

DISCUSSION

Statistical issues. We proposed an approximately calibrated procedure to test the equality of probability vectors p_1 and p_2 of multinomial draws when sample sizes are small and data are sparse. This issue is generic but is especially relevant for microorganisms that are pathogens of plants, as mentioned in the introduction. Based on the simulation study, we give the following practical advice:

- when sample size is small (i.e., a few dozen isolates in the two samples), use the GMCPIC test based on equation 8 that is numerically calibrated and whose power is satisfactory;
- whatever the sample size, when the GMCPIC test based on equation 8 rejects the null hypothesis, the alternative hypothesis is true with the specified significance level;
- fixing the tuning parameters of the test at $B = 10^4$ and $M = 10^3$ led to robust results in terms of test calibration in diverse situations but they can be increased to gain in robustness if computation time is not an issue (Supplementary Text S4);
- simulation studies of type I error and powers can be carried out to determine which tests are calibrated for some given sample sizes and a given number of categories, and to determine which type and which amplitude of discrepancy between p_1 and p_2 can be detected;
- when the type and the amplitude of discrepancies are fixed, the power analysis can help in determining what sample size is required to reject the null hypothesis at a given rate;
- in this article (including Supplementary Figures S3 and S4), we considered sample sizes ranging from 10 to 100 and numbers of categories ranging from 3 to 100. For cases out of these ranges, new simulation studies should be carried out to evaluate the usefulness of the GMCPIC test.

In order to improve the performance of the GMCPIC test, further research should address the choice of the statistic to be calibrated. Cressie and Read (1984, 1989) studied the family of power divergence statistics for testing the fit of observed frequencies to expected frequencies. This family of statistics, including the χ^2 statistic, could be used to define other versions of the GMCPIC test and study whether one of these versions would be more efficient than the GMCPIC test based on the statistic given by equation 8.

Another possible improvement of the test concerns the generalized estimate $\hat{p}(w)$ of the probability vector p under the null hypothesis which, in our procedure, is a convex combination of $\hat{p}_1$ and $\hat{p}_2$ (i.e., $\hat{p}(w) = w\hat{p}_1 + (1-w)\hat{p}_2$, where the weight w is optimized over the interval [0,1] to obtain a calibrated test). To improve the approach, one could search for a generalized estimate (leading to a calibrated test) in a larger space. Allowing w to be larger than 1 or lower than 0 is a possibility but, in our computations, the optimal w most of the time was between 0.25 and 0.95 (Supplementary Fig. S5). Therefore, testing values greater than 1 and lower than 0 for w will generally be a waste of computation time. Allowing $\hat{p}(w)$ to be outside the line joining $\hat{p}_1$ and $\hat{p}_2$ should lead to improving the test calibration; however, this is not a simple issue when the number of categories (or variants) in the vectors of counts is large (because of the curse of dimensionality). This is the main reason why unconditional exact tests have been developed for 2-by-2 contingency tables only. A complementary approach could be to not rely on a single (numerically optimal) value of the weight w (which might produce instability in the test results depending on the case study) but to integrate out the test statistic over w by taking into account a penalization depending on the calibration criterion given in equation 6. Such an approach should be designed in such a way that additional computation cost is negligible.

TABLE 5. P values of the Monte Carlo (MC) χ^2 test with $B = 10^4$, the Fisher's exact test, its MC version with $B = 10^4$, and the generalized Monte Carlo plug-in test with calibration (GMCPIC) test with the statistics of equation 8, $B = 10^4$ and $M = 10^3$ simulations, applied to *Magnaporthe oryzae* compositions sampled in China and Madagascar and to *Pseudomonas syringae* compositions considered at three different resolutions (phylogroups, clades, and haplotypes)

Pathogen	Data set	MC χ^2 test	Fisher's test	MC Fisher's test	GMCPIC test
M. oryzae	China	0.01	<0.0001	0.01	0.01
	Madagascar	0.26	0.34	0.33	0.29
P. syringae	Phylogroup resolution	0.0001	<0.0001	0.0001	<0.0001
	Clade resolution	<0.0001	<0.0001	0.0001	<0.0001
	Haplotype resolution	0.0001	<0.0001	0.0001	<0.0001

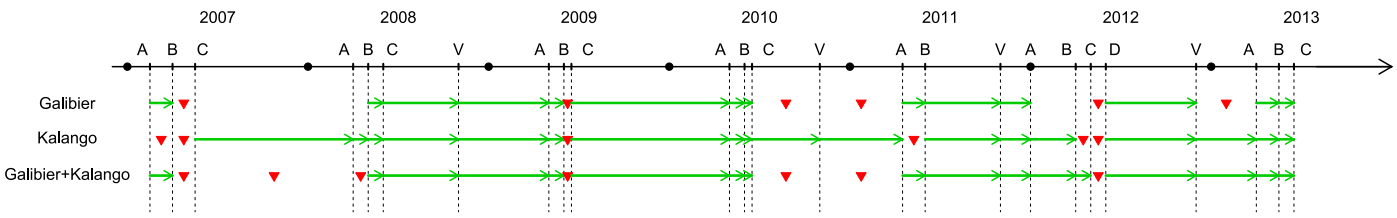

Fig. 5. Results of the tests applied to *Puccinia triticina* data sampled over 7 years (2007 to 2013) on Galibier and Kalango wheat. Arrows indicate equality of vectors of probabilities p_1 and p_2 not rejected, triangles indicate equality rejected, and absence of symbol indicates missing data, implying that no test has been carried out. Tests were separately applied to data collected from Galibier and data collected from Kalango. Tests were also applied to the merged data by taking into account the differences in the virulence (Supplementary Text S3). Letters A, B, C, D, and V refer to different sampling dates in each year of the study period.

In the *P. triticina* case study, the GMCPIC test is applied several times for a temporal series of samples N_1,N_2,N_3,.... Thus, we tackled a multiple-test situation, where the tests were dependent because each sample (except the first and last ones) was used in two tests (N_2 is used when N_1 and N_2 are compared and when N_2 and N_3 are compared). Thus, results for this case study must be carefully interpreted. In this application, we noticed that the null hypothesis is rejected for 25 to 30% of the pairwise comparisons (instead of the expected rate of 5% if the null hypothesis was true during the entire study period and the dependence issue was neglected). Approximately the same percentage of rejections holds when the issue of test dependence is circumvented by comparing only N_1 and N_2, N_3 and N_4, N_5 and N_6, and so on. Therefore, in the *P. triticina* case study, the result of each test cannot be analyzed separately (i.e., specific disruptions in the genetic structure of the local pathogen population cannot be pointed out) but we can draw a conclusion based on the results of all the tests (as we did in the Results section): our analysis does suggest that the local *P. triticina* population experienced a statistically significant number of disruptions during the study period.

Biological issues. *Reproduction of* M. oryzae. In most rice-growing areas (as, for example, in Madagascar), rice blast reproduces clonally (Saleh et al. 2012) by producing asexual spores. Epidemics probably start from infected seed, which produce spores that infect leaves and produce mycelium. After 5 to 7 days, lesions appear that will produce asexual spores under favorable conditions. Young plants are particularly susceptible and are heavily infected. With aging, rice acquires a so-called adult resistance, making infection by the pathogen more difficult. During the emergence (heading) of the rice inflorescence (panicle), the last leaf (flag leaf) is highly susceptible to the blast pathogen and favors panicle infection. Because the physiology of the leaves and the panicle are very different, and because of inconsistent published results on pathotype composition, whether populations sampled on the two types or organs during the same epidemic are identical is controversial. The GMCPIC test developed in this study was applied to two populations sampled in Madagascar in the same field on leaves and panicles at the beginning and the end of the growing season, respectively. The equality of PC was not rejected, confirming that the reproduction was strictly clonal and that there was no bottleneck and demonstrating that the genetic composition of the population was not different between the two sampling stages.

In Yunnan Province of China, the putative center of origin of *M. oryzae* on rice (Saleh et al. 2014), we previously demonstrated that sexual reproduction is taking place in at least one population (Saleh et al. 2012). In this case, generalized recombination is expected to shuffle alleles at different loci and create new and unique MLG. Here, we confirmed that the PC in terms of MLG was different. In both situations, the result of the test matched the expectations. Therefore, the *M. oryzae* case clearly validates our procedure for comparing PC.

Spatial structure of Pseudomonas syringae *populations. Pseudomonas syringae* designates a complex of plant-pathogenic bacteria associated with numerous past and present diseases across the world. Phylogroups and clades of the *P. syringae* complex are phenotypically diverse, and no distinct ecology can be attributed to most of them (Berge et al. 2014). This diversity is found both on pathogen populations collected from cultivated plants and from their saprophytic relatives collected in different environments of the water cycle, such as leaf litter, streams, snow, or wild plants in alpine areas (Morris et al. 2013). All these habitats contribute to the evolution and emergence of new pathotypes by exerting selection pressures on determinants associated with pathogenicity. Aerial transport is a means of dissemination of these pathotypes and precipitation may lead to the deposition of these populations in new areas. Comparing genetic patterns of diversity in precipitation of two very contrasted habitats may provide clues about local adaptation and population mixing between these habitats.

Genotyping of highly conserved core genome genes is a reliable approach to assess the diversity of *P. syringae*, which is represented by 13 phylogroups and 26 clades (Berge et al. 2014). Only genotyping of the *cts* gene is discriminatory enough to address *P. syringae* diversity at a satisfying resolution (Berge et al. 2014). However, its cost limits sequencing efforts to a very few strains per sample. Therefore, contexts such as this one, where there are data for only a few strains per sample, are likely to be a frequent limiting factor when analyzing population structure based on gene sequences. We maximized the diversity of precipitation events sampled within ecosystems to better approximate the real population structure of emission sources in these ecosystems. Each comparison made at a specific resolution (haplotype, clade, or phylogroup) gave access to a different level of diversity and all rejected the null hypothesis. Therefore, composition of populations in precipitation is different according to the area (UDR versus LDR). Some phylogroups, clades, or haplotypes are present in both areas (e.g., phylogroup 2), whereas others are absent from one or the other region. Importantly, the dominant groups (e.g., phylogroup 10) are different in each ecosystem. Overall, we formally demonstrated that each ecosystem is associated with different *P. syringae* populations. These results corroborate the hypothesis that (i) different land occupation and fragmentation of landscapes structure plant pathogen populations and (ii) groups within the *P. syringae* complex may effectively have different ecologies. Implications for epidemiology are important because they suggest that dissemination of emerging or reemerging pathotypes may be fostered by land management. Furthermore, a previous study of the biogeography of *P. syringae* did not reveal differences in population structure for different geographic locations in spite of a high frequency of endemic haplotypes (Morris et al. 2010), suggesting the possible lack of sufficient statistical power of the population genetic analyses used in this previous study.

Temporal recurrence of Puccinia triticina. Disruptions in PC appeared to be more frequent within a cropping season (one-third of the cases) than during the intercrop period (over four intercrop periods with data collected on volunteers, two disruptions were detected for only one of the cultivars [Galibier]). Therefore, the intercrop period does not represent a major bottleneck for the population dynamic of the fungus. This is consistent with the generally admitted view that wheat volunteers serve as a "green bridge" allowing the survival of the fungus during the intercrop period (Moschini and Pérez 1999; Singh et al. 2004). Wheat volunteers represent the only hosts widely available to the fungus after harvest. The strong clonal structure of the local populations of the pathogen (Goyeau et al. 2007) indicates that sexual reproduction on an alternate host, were it observed in the area of study, would be of little practical significance; the only wild grass the pathogen could infect, *Aegilops ovata* (Dupias 1952), has not been recently recorded in local botanical surveys (Tela Botanica network, Montpellier, France; http://www.tela-botanica.org/bdtfx-nn-957).

Disruptions in PC during the cropping season were not expected, because the increase of disease during the season is generally believed to be caused by local multiplication of the pathogen. The huge sporulation capacity of the fungus and the swift progress of the epidemic are expected to provide demographic advantage to the local pathogen population. Thus, it is likely that populations windblown from neighboring plots during the course of the epidemic modified the population structure in our observation plots. Moreover, infection by wind-dispersed spores of remote origin cannot be firmly excluded. In Europe, two proposed pathways for the spread of stem rust, caused by another rust fungus (*Puccinia graminis* f. sp. *tritici*), are partially supported by empirical evidence; in contrast, there is no hint of a regular continental spread of wheat leaf rust, caused by *P. triticina* (Zadoks and Bouwman 1985).

Concluding remarks. Today, emphasis is legitimately put by plant pathologists on accelerating exploitation of big data (Saunders 2015). In contrast, some generic questions are intrinsically connected to small samples and sparse data sets. Comparing the genetic composition of small-sized populations of microorganisms

is one such classical but difficult issue. The GMCPIC test developed in this study provides a robust alternative to routine tests, which have well-known limits (or limits that should be known) when applied to small samples. We illustrated the power of the GMCPIC test on three case studies in plant disease epidemiology where we considered the big data approach to be unmanageable in practice. We expect the GMCPIC test to be used by the whole community of plant pathologists and, hopefully by other biologists addressing the same kinds of issues (e.g., geneticists and ecologists).

ACKNOWLEDGMENTS

We thank G. Lagarde and the staff of the local supply cooperative Qualisol for their great contribution to *P. triticina* sampling in the region of Lomagne (France); J. F. Rey from INRA for the implementation of the computer code into an R package; and P. Xu (YAAS, China), Dodelys Andrianatsimialona (FOFIFA, Madagascar) for sharing *M. oryzae* strains or permitting collection of samples. This research was funded by the European Union Seventh Framework Programme (PLANTFOODSEC, 261752). *M. oryzae* genotypic data used in this work were produced through molecular genetic analysis technical facilities of the labex "Centre Méditerranéen de l'Environnement et de la Biodiversité". We thank CIRAD, INRA, Agropolis Fondation ("Rice blast networking" project), and ANR (Emerfundis ANR-Biodiv-07) for financial support of the work on *M. oryzae*.

LITERATURE CITED

Agresti, A. 2007. An Introduction to Categorical Data Analysis. John Wiley & Sons Ltd., Chichester, UK.

Arnaud-Haond, S., Duarte, C. M., Alberto, F., and Serrão, E. A. 2007. Standardizing methods to address clonality in population studies. Mol. Ecol. 16: 5115-5139.

Berge, O., Monteil, C. L., Bartoli, C., Chandeysson, C., Guilbaud, C., Sands, D. C., and Morris, C. E. 2014. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. PLoS One 9:e105547.

Burdon, J. J. 1993. The structure of pathogen populations in natural plant communities. Annu. Rev. Phytopathol. 31:305-323.

Cressie, N., and Read, T. R. C. 1984. Multinomial goodness-of-fit tests. J. R. Stat. Soc. B Met. 46:440-464.

Cressie, N., and Read, T. R. C. 1989. Pearson's X^2 and the loglikelihood ratio statistics G^2. A comparative review. Int. Stat. Rev. 57:19-43.

Dupias, G. 1952. À propos des Urédinées parasites des Ægilops. Ber. Schweiz. Bot. Ges. 62:370-373.

García-Arenal, F., Fraile, A., and Malpica, J. M. 2001. Variability and genetic structure of plant viral populations. Annu. Rev. Phytopathol. 39:157-186.

Gérard, P. R., Husson, C., Pinon, J., and Frey, P. 2006. Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. Phytopathology 96:1027-1036.

Gilbert, G. S. 2002. Evolutionary ecology of plant diseases in natural ecosystems. Annu. Rev. Phytopathol. 40:13-43.

Goyeau, H., Berder, J., Czerepak, C., Gautier, A., Lanen, C., and Lannou, C. 2012. Low diversity and fast evolution in the population of *Puccinia triticina* causing durum wheat leaf rust in France from 1999 to 2009, as revealed by an adapted differential set. Plant Pathol. 61:761-772.

Goyeau, H., Halkett, F., Zapater, M. F., Carlier, J., and Lannou, C. 2007. Clonality and host selection in the wheat pathogen fungus *Puccinia triticina*. Fungal Genet. Biol. 44:474-483.

Goyeau, H., Park, R., Schaeffer, B., and Lannou, C. 2006. Distribution of pathotypes with regard to host cultivars in French wheat leaf rust populations. Phytopathology 96:264-273.

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. B 30:582-598.

Hull, R. 2008. The phenotypic expression of a genotype: Bringing muddy boots and micropipettes together. Annu. Rev. Phytopathol. 46:1-11.

Kolmer, J. A., Hanzalova, A., Goyeau, H., Bayles, R., and Morgounov, A. 2012. Genetic differentiation of the wheat leaf rust fungus *Puccinia triticina* in Europe. Plant Pathol. 62:21-31.

Leroy, T., Le Cam, B., and Lemaire, C. 2014. When virulence originates from non-agricultural hosts: New insights into plant breeding. Infect. Genet. Evol. 27:521-529.

Linde, C. C., Zhan, J., and McDonald, B. A. 2002. Population structure of *Mycosphaerella graminicola*: From lesions to continents. Phytopathology 92:946-955.

Lydersen, S., Fagerland, M. W., and Laake, P. 2009. Recommended tests for association in 2x2 tables. Stat. Med. 28:1159-1175.

Manly, B. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman and Hall, London.

McDonald, B. A., and Linde, C. 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu. Rev. Phytopathol. 40: 349-379.

Mehrotra, D. V., Chan, I. S. F., and Berger, R. L. 2003. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics 59:441-450.

Milgroom, M. G. 2015. Population Biology of Plant Pathogens. American Phytopathological Society Press, St. Paul, MN.

Monteil, C. L., Bardin, M., and Morris, C. E. 2014a. Features of air masses associated with the deposition of *Pseudomonas syringae* and *Botrytis cinerea* by rain and snowfall. ISME J. 8:2290-2304.

Monteil, C. L., Lafolie, F., Laurent, J., Clement, J. C., Simler, R., Travi, Y., and Morris, C. E. 2014b. Soil water flow is a source of the plant pathogen *Pseudomonas syringae* in subalpine headwaters. Environ. Microbiol. 16:2038-2052.

Morris, C. E., Monteil, C. L., and Berge, O. 2013. The life history of *Pseudomonas syringae*: Linking agriculture to Earth system processes. Annu. Rev. Phytopathol. 51:85-104.

Morris, C. E., Sands, D. C., Vanneste, J. L., Montarry, J., Oakley, B., Guilbaud, C., and Glaux, C. 2010. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. MBio 1:e00107-10.

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffière, A., Yan, S., Dominguez, H., and Thompson, B. M. 2008. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISME J. 2:321-334.

Moschini, R. C., and Pérez, B. A. 1999. Predicting wheat leaf rust severity using planting date, genetic resistance, and weather variables. Plant Dis. 83: 381-384.

Pule, B. B., Meitz, J. C., Thompson, A. H., Linde, C. C., Fry, W. E., Langenhoven, S. D., Meyers, K. L., Kandolo, D. S., van Rij, N. C., and McLeod, A. 2013. *Phytophthora infestans* populations in central, eastern and southern African countries consist of two major clonal lineages. Plant Pathol. 62:154-165.

Rozenfeld, A. F., Arnaud-Haond, S., Hernandez-Garcia, E., Eguiluz, V. M., Matias, M. A., Serrao, E., and Duarte, C. M. 2007. Spectrum of genetic diversity and networks of clonal organisms. J. R. Soc. Interface 4:1093-1102.

Saleh, D., Milazzo, J., Adreit, H., Fournier, E., and Tharreau, D. 2014. South-East Asia is the center of origin, diversity and dispersion of the rice blast fungus, *Magnaporthe oryzae*. New Phytol. 201:1440-1456.

Saleh, D., Xu, P., Shen, Y., Li, G. C., Adreit, H., Milazzo, J., Ravigne, V., Bazin, E., Notteghem, J. L., Fournier, E., and Tharreau, D. 2012. Sex at the origin: An Asian population of the rice blast fungus *Magnaporthe oryzae* reproduces sexually. Mol. Ecol. 21:1330-1344.

Saunders, D. G. O. 2015. Hitchhiker's guide to multi-dimensional plant pathology. New Phytol. 205:1028-1033.

Sellke, T., Bayarri, M. J., and Berger, J. O. 2001. Calibration of p values for testing precise null hypotheses. Am. Stat. 55:62-71.

Singh, R. P., Huerta-Espino, J., Pfeiffer, W., and Figueroa-Lopez, P. 2004. Occurrence and impact of a new leaf rust race on durum wheat in northwestern Mexico from 2001 to 2003. Plant Dis. 88:703-708.

Tian, Y., Yin, J., Sun, J., Ma, H., Quan, J., and Shan, W. 2015. Population structure of the late blight pathogen *Phytophthora infestans* in a potato germplasm nursery in two consecutive years. Phytopathology 105:771-777.

Tollenaere, C., and Laine, A. L. 2013. Investigating the production of sexual resting structures in a plant pathogen reveals unexpected self-fertility and genotype-by-environment effects. J. Evol. Biol. 26:1716-1726.

Villaréal, L. M. M. A., and Lannou, C. 2000. Selection for increased spore efficacy by host genetic background in a wheat powdery mildew population. Phytopathology 90:1300-1306.

Vinatzer, B. A., Monteil, C. L., and Clarke, C. R. 2014. Harnessing population genomics to understand how bacterial pathogens emerge, adapt to crop hosts, and disseminate. Annu. Rev. Phytopathol. 52:19-43.

Xu, J., ed. 2010. Microbial Population Genetics. Calster Academic Press, Norfolk, UK.

Zadoks, J. C., and Bouwman, J. J. 1985. Epidemiology in Europe. Pages 329-369 in: The Cereal Rusts: Vol. II. Disease, Distribution, Epidemiology and Control. A. P. Roelfs and W. R. Bushnell, eds. Academic Press, Orlando, FL.