Testing differences between pathogen compositions with small samples and sparse data

Samuel Soubeyrand[1*], Vincent Garreta[12], Caroline Monteil[3], Frédéric Suffert[2], Henriette Goyeau[2], Julie Berder[2], Jacques Moinard[4], Elisabeth Fournier[5], Didier Tharreau[6], Cindy E. Morris[3], Ivan Sache[7]

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

[2]INRA, UMR1290 Bioger, AgroParisTech, Université Paris-Saclay 78850 Thiverval-Grignon, France

[3]INRA, UR0407 Plant Pathology, 84143 Montfavet, France

[4]DRAAF Midi-Pyrénées, 31074 Toulouse Cedex, France

[5]INRA, UMR BGPI, 34398 Montpellier, France

[6]CIRAD, UMR BGPI, 34398 Montpellier, France

[7]AgroParisTech, UMR1290 Bioger, 78850 Thiverval-Grignon, France

[*] Corresponding author: samuel.soubeyrand@inra.fr

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 if 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, namely the Generalized Monte Carlo Plug-In test with Calibration (GMCPIC test), which is implemented in an R package available at http://dx.doi.org/10.5281/zenodo.53996. 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*.

Introduction

The genetic structure of pathogen populations is an important driver of epidemics affecting crops (Garcia-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 (Villareal 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 (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 pre-requisite, while several pathogens do not show any signature of such a reproduction and exhibit clonal population structure; the sexual system is, moreover, 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. Besides, 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 sample sizes are small, data are also sparse for non-dominant variants. Therefore, statistical tests adapted to small sample sizes and sparse data are needed.

For comparing vectors of proportions, the Chi-squared test and the Fisher's exact test (Agresti 2007, chap. 2-3) 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, chap. 12) 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 2x2 contingency tables (i.e. with 2 groups and 2 response types); see reviews by Mehrotra et al. (2003) and Lydersen et al. (2009). These tests rely on a maximization with respect to the probabilities of all the response types. For a 2x2 contingency table, there are 2 response types with probabilities π and 1-π under the null hypothesis and, therefore, the maximization is carried out with respect to a single parameter, namely π. This maximization approach adopted in unconditional exact tests prevents their application for larger contingency tables. Indeed, if we consider a design with 2 groups and n response types (like 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 dozens of variants of the pathogen including a large number of non-dominant variants, and a few dozens of isolates), without any a priori biological knowledge. This procedure has a wide spectrum of applications since, 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.e. (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 R package GMCPIC (Generalized Monte Carlo Plug-In test with Calibration) and available at http://dx.doi.org/10.5281/zenodo.53996.

Material 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 Chi-squared test and the Fisher's exact tests 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 Chi-squared statistics Q used in the Chi-squared test:

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

where K is the number of categories, N_{ij} is the j-th component of N_i, $j \in \{1, \dots, K\}$ 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:

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

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}, i.e. the ratios $n_i \hat{q}_j / E(N_{ij})$ vary strongly. Therefore, when the

classical Chi-squared test is applied, the normal approximation of the distribution of $n_i\hat{q}_j$ is crude and so is the Chi-squared approximation of the distribution of the statistic Q. Similarly, when the Monte Carlo version of the Chi-squared 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 (definitions: 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); 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 2x2 contingency tables, unconditional exact tests have been proposed to solve the issue mentioned above (Mehrotra et al. 2003; Lydersen et al. 2009) 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.

Generalized Monte Carlo plug-in test with calibration (GMCPIC test)

As seen above, with sparse data, the inadequacy of the Chi-squared 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, $N_1 \sim \text{Multinomial}(n_1, p_1)$, $N_2 \sim \text{Multinomial}(n_2, p_2)$ and $p = p_1 = p_2$). For example, in the Chi-squared 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 $N_2^{(b)}$ ($b \in \{1, \dots, B\}$, 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 $S\left(N_1^{(b)}, N_2^{(b)}, w\right)$ less than or equal to $S(N_1, N_2, w)$:

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

  where $I\{E\} = 1$ if event E occurs, zero otherwise.

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

- independently draw $2M$ samples $\widetilde{N}_1^{(m)}$ and $\widetilde{N}_2^{(m)}$ ($m \in \{1, \dots, M\}$, 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:

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

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

The GMCPIC test can be applied with various statistics, especially the extension of the negative Chi-squared statistic:

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

and the extension of the statistic used in the plug-in test without calibration (see Supplementary Text S1):

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

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 $\hat{p}(\widetilde{w})$ defined as a weighted mean of the observed probability vectors $\hat{p}_1$ and $\hat{p}_2$, but (ii) the estimate $\hat{p}(\widetilde{w})$ is obtained via a numerical maximization, like in unconditional tests. However, in contrast with 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 as the test is calibrated or, in other words, such as 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 "$N_1 \sim \text{Multinomial}(n_1, p)$, $N_2 \sim \text{Multinomial}(n_2, p)$, $N_1 \perp N_2$", 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 $\widetilde{w}$ of Equation (6) is not in general equal to $\frac{n_1}{n_1+n_2}$, 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 1000) 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, 1/3) or heterogeneous

probabilities (0.80, 0.19, 0.01). For vectors with 33 categories, we used heterogeneous probabilities (0.70, 0.10, 0.10, 0.10/30,..., 0.10/30). One thousand 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 if 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 in this aim. To complete these series, we provide a more generic simulation study where vectors of probabilities are randomly generated with varying means and variances (Supplementary Table S1).

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 (higher the rejection rate, 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 1000). 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; see 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. One thousand 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 (thereafter, PC stands for pathogen compositions). 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 genoptype (*M. oryzae*), a virulence phenotype (*P. triticina*), a haplotype, a clade or a phylogroup (*P. syringae*). For highly-diverse pathogens, the number of different variants that are considered may be large, and zeros in vector describing the pathogen composition 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.

Magnaporthe oryzae

Two data sets were collected in Madagascar and China for studying the reproduction regime of a rice fungal pathogen, *Magnaporthe 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 Fig. S2. Field samples were purified by sub-culturing 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. The pathogen compositions

(summarized in Table 1; detailed in Fig. 1) therefore consist in frequencies of MLG in each population.

This application was selected as a means to validate our procedure: we expect that the similarity of pathogen compositions 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 multilocus genotypes 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 re-assortment of allelic associations, thus creating new multilocus genotypes. Frequencies of multilocus genotypes thus vary between two samples separated by at least one event of sexual reproduction.

Pseudomonas 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 (see map in Supplementary Fig. S2), to investigate the spatial diversity of *P. syringae* populations (Monteil et al. 2014). Both environments may be connected since members of *P. syringae* strains are able to disseminate through air and water fluxes (Monteil et al. 2014a,b). The pathogen compositions taken from the Alps and the pathogen compositions taken from the crops were considered at three different resolutions, namely haplotypes, clades and phylogroups (Berge et al. 2014). Our testing procedure was applied to pathogen compositions 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 Low 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 pathogen compositions 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). The pathogen compositions (summarized in Table 2; detailed in Fig. 2) therefore consisted of frequencies of haplotypes, clades and phylogroups in the upper and lower basins of the Durance.

Puccinia triticina

The temporal recurrence of *Puccinia triticina* was investigated by collecting leaves infected with *P. triticina* for seven consecutive years (2007-2013) from wheat fields located in South-West France (see map in Supplementary Fig. S2). 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 were 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 (a) temporal disruptions and continuations in the genetic structure of the local pathogen population and, more specifically, (b) the role of wheat volunteers in the yearly recurrence of disease in wheat crops. The null hypothesis was that over-summering of the pathogen on volunteers led to local perpetuation of disease over the whole period of study; accordingly, a single, multi-year epidemic would have occurred rather than successive yearly epidemics re-initiated 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 *P. triticina*. Infection types on the differentials were evaluated 10 days after inoculation to establish the virulence phenotype of each strain. The

pathogen compositions (summarized in Table 3; detailed in Fig. 3) therefore consisted of frequencies of virulence phenotypes at each sampling date.

Results

Simulation-based study: analysis of type I error and power

Table 4 gives assessments of type I errors (i.e. the incorrect rejection of the true null hypothesis) in different settings. The Chi-squared 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 Chi-squared test and the two Fishers tests lead to very close type I errors. The Monte Carlo plug-in test consistently over-rejects the null hypothesis in any settings and is definitely an incorrect test. The GMCPIC test based on Eq. (7) shows an incorrect type I error for the large pathogen compositions at small and moderate sample sizes (trend to under-rejection), whereas the GMCPIC test based on Eq. (8), which is a calibrated version of the Monte Carlo plug-in test, has correct type I errors in every settings. 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 PCs with 33 variants, with varying type of difference between the two PCs, and with varying amplitude δ of the difference, are compared (Fig. 4). First, it has to be noted that the Monte Carlo Chi-squared test and both Fisher tests have similar powers in all settings (however, the Fisher's exact test was not run for samples with size 1000 because of excessive computation time). Second, the GMCPIC test based on Eq. (8) (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 Eq. (8) can be lower than the power of the Monte Carlo Chi-squared test and the two Fisher 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 provided in Supplementary Table S1, where the vectors of probabilities p_1 and p_2 are randomly generated with varying means and variances.

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

Reproduction of *Magnaporthe oryzae*

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

The similarity of PCs 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 PCs 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 Eq. (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 *P. 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 PCs 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 *P. triticina* data, the tests were applied to each pair of consecutive samples collected only in fields sown with Galibier and to each pair of consecutive samples collected only in fields sown with Kalango. 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.

Fig. 5 shows at which periods the temporal continuation in the genetic structure of the local *P. triticina* population (i.e. the null hypothesis) is rejected by the GMCPIC test based on Eq. (8), that is to say when there are disruptions in the pathogen composition (Supplementary Table S2, provides the corresponding *p*-values). The total continuation of the epidemic with constant composition over the study period (2007-2013) is rejected for both cultivars Galibier and Kalango. Indeed, for Galibier (resp. Kalango) 30% (resp. 25%) of the tests 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 re-infect the wheat crops).

Supplementary Table S3, compares the results obtained with the GMCPIC test based on Eq. (8), the Monte Carlo Chi-square test and both Fisher tests. 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 low 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 dozens of isolates in the two samples), use the GMCPIC test based on Eq. (8) that is numerically calibrated and whose power is satisfactory;
- Whatever the sample size, when the GMCPIC test based on Eq. (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 lead 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 (see 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 Table S1, Supplementary Text S4, and 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 chi-squared statistic, could be used to define other versions of the GMCPIC test, and study if 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 is in our procedure 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 was most of the time between 0.25 and 0.95; See Supplementary Text S4 and Supplementary Figures S3, S4 and 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 improve the test calibration, but 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 2x2 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.

In the *P. triticina* case study, the GMCPIC test is applied several times for a temporal series of samples $N_1, N_2, N_3, \ldots$. Thus, we tackle a multiple test situation, where the tests are dependent because each sample (except the first and last ones) is 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 25-30% of the pairwise comparisons (instead of the expected rate 5% if the null hypothesis was true during all the study period and the dependence issue is 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 analysed 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 result 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 is reproducing clonally (Saleh et al. 2012) by producing asexual spores. Epidemics probably start from infected seeds which produce spores that infect leaves and produce mycelium. After 5-7 days lesions appear, that will produce asexual spores under favourable conditions. Young plants are particularly susceptible and are heavily infected. With aging rice is acquiring a so called adult resistance making infection more difficult by the pathogen. During the emergence (heading) of the rice inflorescence (panicle), the last leaf (flag leaf) is highly susceptible to the blast pathogen and favours panicle infection. Since 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 on 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 PCs was not rejected, confirming that the reproduction is strictly clonal and that there was no bottleneck and demonstrating that the genetic composition of the population is not different between the two sampling stages.

In Yunnan Province of China, the putative centre 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 multilocus genotypes. Here we confirmed that the PC in terms of multilocus genotypes was different. In both situations, the result of the

test matches the expectations. The *M. oryzae* case therefore clearly validates our procedure for comparing PCs.

Spatial structure of *P. 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 to 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 core genome genes highly conserved is a reliable approach to assess the diversity of *P. syringae* which is represented by 13 phylogroups and 26 clades (Berge et al. 2014). Genotyping of the *cts* gene only is discriminatory enough to address *P. syringae* diversity at a satisfying resolution (Berge et al. 2014). However, its cost limits sequencing effort 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 vs. LDR). Some phylogroups, clades or haplotypes are present in both areas (e.g., phylogroup 2), while 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 pathogens populations and (ii) groups within the *P. syringae* complex may effectively have different ecologies.

Implications for epidemiology are important because it suggests 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. 2012) suggesting the possible lack of sufficient statistical power of the population genetic analyses used in this previous study.

Temporal recurrence of *P. triticina*

Disruptions in pathogen compositions 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, namely 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 (Moschini and Perez 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, would it be observed in the area of study, would be of little practical significance; the only wild grass the pathogen could infect, *Ægilops 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 pathogen compositions during the cropping season were not expected, since 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. It is thus likely that populations wind-blown from neighbouring 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 (*P. graminis f.sp. tritici*) are partially supported by empirical evidence; contrastively, 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 (Saunder, 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 micro-organisms is such a 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 consider the big data approach as not manageable 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.

Acknowledgements

We thank G. Lagarde and staff of the local supply cooperative Qualisol for their great contribution to *P. triticina* sampling in the region of Lomagne (France). We thank J.F. Rey from INRA for the implementation of the computer code into an R package. We are grateful to Peng 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). *Magnaporthe 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 to the work on *M. oryzae*.

Literature Cited

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

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. Roy. 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. Berichte der Schweizerischen Botanischen Gesellschaft 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. Journal of the Royal Statistical Society 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. & 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. APS Press, St. Paul, USA.

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., 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., 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., 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., Duarte, C.M. 2007. Spectrum of genetic diversity and networks of clonal organisms. J. Roy. Soc. Interface 4: 1093-1102.

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

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.

Saunders, D.G.O. 2015. Hitchhiker's guide to multi-dimensional plant pathology. New Phytologist 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. & 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., 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. Evolution. 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.

Tables

Table 1. Number of *Magnaporthe oryzae* isolates and number of different variants in each sample. For *M. oryzae* sampled in China: 1st sample collected in August 2008; 2nd sample collected in September 2009. For *M. oryzae* sampled in Madagascar: 1st sample collected in February 2005; 2nd sample collected in April 2005.

Statistic	Data set	1st sample	2nd sample	Pooled
Number of isolates	China	24	83	107
	Madagascar	17	40	57
Number of different variants	China	21	72	92
	Madagascar	10	12	18

Table 2. Number of *Pseudomonas syringae* isolates and number of different variants in each sample. For *P. syringae:* 1st sample collected in UDR basin (alpine samples); 2nd sample collected in LDR basin (agriculture samples); variants are defined with respect to three different resolutions: phylogroups, clades and haplotypes.

Statistic	Resolution	1st sample	2nd sample	Pooled
Number of isolates	All resolutions	100	110	210
Number of different variants	Phylogroups resolution	5	8	10
	Clade resolution	9	14	17
	Haplotype resolution	19	44	57

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

Year	2007				2008				2009				2010			
Sampling #	A	B	C	V[a]	A	B	C	V	A	B	C	V	A	B	C	V
Sampling date	Feb. 22	Apr. 04	May 14		Apr. 01	May 05	May 26	Oct. 28	May 05	May 25	June 10	Oct. 27	May 04	May 26	June 15	Oct. 26
Number of isolates for which the pathotype was determined																
Galibier	24	11	24		0[b]	24	28	10	17	30	27	0[b]	11	29	30	5
Kalango	19	17	22		28	18	30	10	13	28	29	0[b]	10	27	30	10
Total	43	28	46		28	42	58	20	30	58	56	0[b]	21	56	60	15
Number of different pathotypes																
Galibier	9	6	9		0	4	6	4	8	1	5	0	4	9	9	2
Kalango	6	6	10		8	8	8	3	5	8	7	0	4	11	12	6
Both	12	10	14		8	10	13	6	11	9	11	0	6	16	18	7

Year	2011				2012					2013			Pooled
Sampling #	A	B	C[c]	V	A	B	C	D	V	A	B	C	
Sampling date	Apr. 21	May 27		Oct. 26	Dec. 29	Mar. 29	Apr. 25	May 24	Nov. 6	Mar. 24	May 17	June 13	
Number of isolates for which the pathotype was determined													
Galibier	23	29		10	21	0[b]	10	30	11	30	30	30	494
Kalango	28	29		10	27	8	8	26	13	30	30	30	530
Total	51	58		20	48	8	18	56	24	60	60	60	1024
Number of different pathotypes													
Galibier	10	11		7	6	0	4	13	8	14	12	9	50
Kalango	6	8		3	3	2	4	5	5	11	7	8	37
Both	12	13		8	7	2	6	14	11	19	14	12	64

[a] No sampling on volunteers.
[b] No infected leaf was found.
[c] No third sampling on cvs. Galibier and Kalango in 2011.

Table 4. Type I errors of the chi-squared test and its Monte Carlo version (with B=10^4 simulations), the Fisher's exact test and its Monte Carlo version (with B=10^4 simulations), the Monte Carlo plug-in test (with B=10^4 simulations), and the GMCPIC test using the statistic of Eq. (7) or the statistic of Eq. (8) and using B=10^4 and M=10^3 simulations. Type I error was computed, for each type of pathogen composition and each sample size, as the proportion of rejections over 1000 repetitions. Between brackets: p-values of the test of equality of the type I errors to the value 0.05. For 33 variants and samples of size 1000, the Fisher's exact test was not run because of excessive computation time.

Pathogen composition	Sample size	χ^2 test	MC χ^2 test	Fisher test	MC Fisher test	MC plug-in test	GMCPIC test with stat. of Eq. (7)	GMCPIC test with stat. of Eq. (8)
3 variants with homogeneous probabilities	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)
	1000	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)
3 variants with heterogeneous probabilities	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)
	1000	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)
33 variants with heterogeneous probabilities	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)
	1000	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)

Table 5. P-values of the Monte Carlo Chi-squared test with B=10^4, the Fisher's exact test, its Monte Carlo version with B=10^4, and the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3 simulations, applied to *M. oryzae* compositions sampled in China and Madagascar and to *P. syringae* compositions considered at three different resolutions, namely phylogroups, clades and haplotypes.

Pathogen	Data set	MC χ^2 test	Fisher test	MC Fisher test	GMCPIC test based on Eq. (8)
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

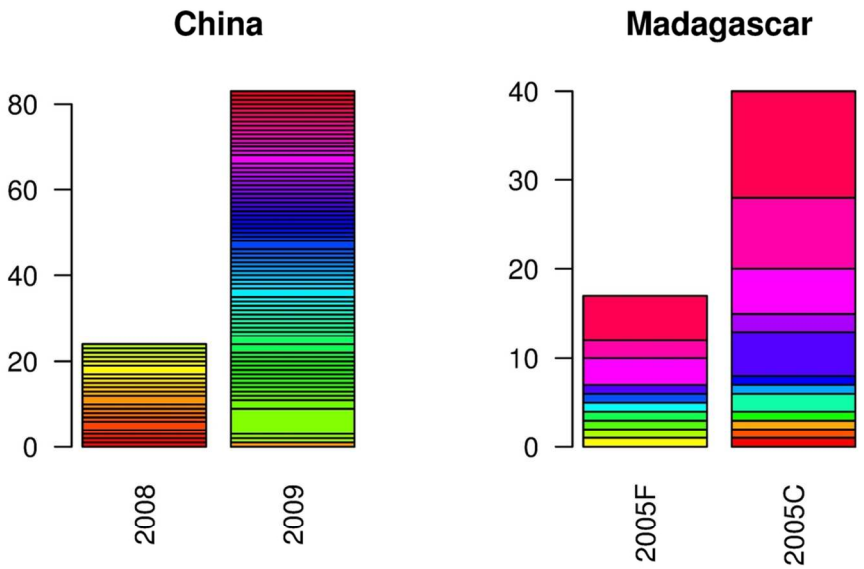


Figure 1. Compositions of populations of Magnaporthe oryzae corresponding to samples collected in China (1st sample collected in August 2008; 2nd sample collected in September 2009) and Madagascar (1st sample collected in February 2005; 2nd sample collected in April 2005). Each colored layer corresponds to a given variant; the height of each layer is proportional to the number of isolates from the corresponding variant.

101x67mm (300 x 300 DPI)

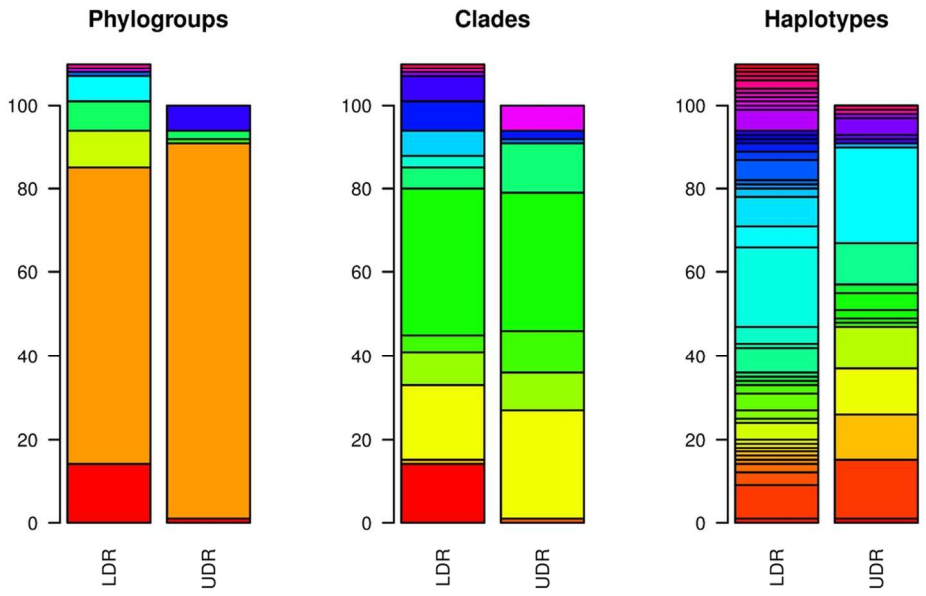


Figure 2. Compositions of populations of Pseudomonas syringae corresponding to samples collected in Upper Durance River (UDR) basin and in Lower Durance River (LDR) basin considered at three different resolutions (phylogroups, clades and haplotypes). Each colored layer corresponds to a given variant;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

101x67mm (300 x 300 DPI)

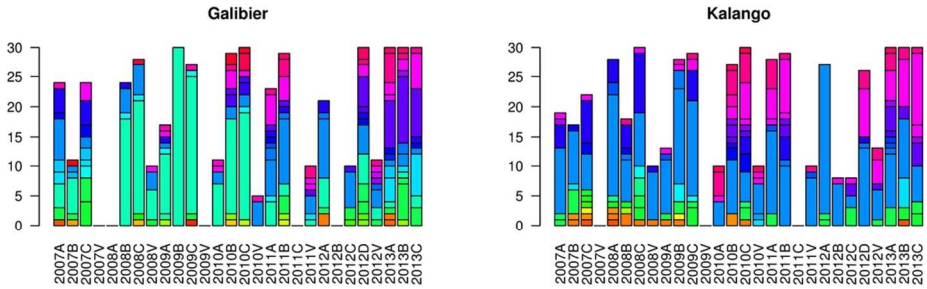


Figure 3. Compositions of populations of Puccinia triticina sampled across time in Galibier and Kalango crops in Southwestern France from 2007 to 2013. Letters A, B, C, D and V refer to different sampling dates on each year of the study period; see Table 3. Each colored layer corresponds to a given variant;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

101x33mm (300 x 300 DPI)

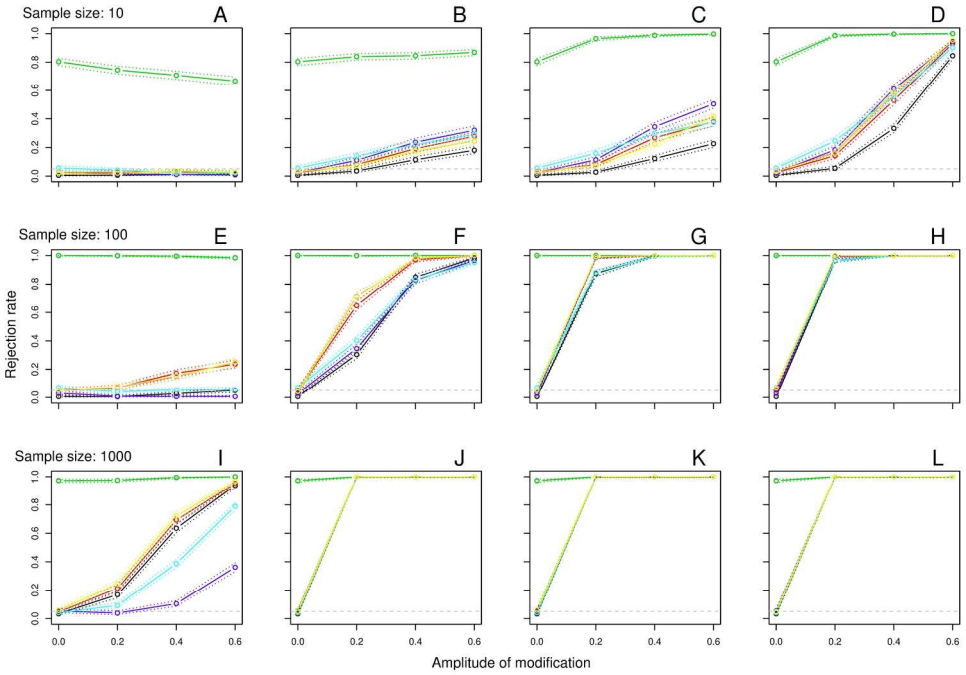


Figure 4. Variation in the powers of the tests with respect to the difference between the vectors of probabilities p1 and p2 (the difference between p1 and p2 depends on the modification type and the amplitude δ of the modification). Each panel corresponds to a specific sample size and a specific modification type. In each panel, the colored curves give the variation in the powers of the following tests: chi-squared test (black) and its Monte Carlo version with B=104 simulations (red), the Fisher's exact test (pink) and its Monte Carlo version with B=104 simulations (yellow), the Monte Carlo plug-in test with B=104 simulations (green), the GMCPIC test using the statistic of Eq. (7) (blue) and the statistic of Eq. (8) (turquoise), with B=104 and M=103 simulations. The powers were assessed over 1000 repetitions for each modification type and each sample size. In all panels, p1=(0.70,0.10,0.10,0.10/30,...,0.10/30). In panels A, E, I (modification type 1): p2=(0.70+δ,0.10,0.10,0.10/30,...,0.10/30)/(1+ δ); in panels B, F, J (modification type 2): p2= (0.70,0.10+δ,0.10,0.10/30,...,0.10/30)/(1+ δ); in panels C, G, K (modification type 3): p2= (0.70,0.10,0.10,0.10/30+δ,...,0.10/30)/(1+ δ); and in panels D, H, L (modification type 4): p2= (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 the type I error provided in Table 4. In each panel, the horizontal dashed grey 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 1000 (bottom panels), the Fisher's exact test was not run because of excessive computation time.

209x148mm (300 x 300 DPI)

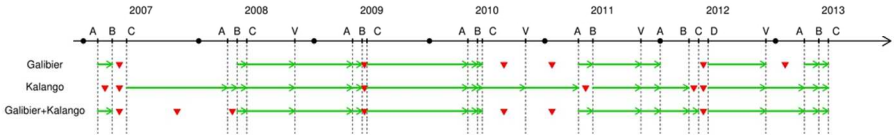


Figure 5. Results of the tests applied to Puccinia triticina data sampled over seven years (2007-2013) on Galibier and Kalango wheat cultivars. Arrows: equality of vectors of probabilities p1 and p2 not rejected; Triangles: equality rejected; Absence of symbol: missing data implying that no test has been carried out. The tests were separately applied to data collected over the Galibier cultivar and data collected over the Kalango cultivar. The tests were also applied to the merged data by taking into account the differences in the virulences (see Supplementary Text S3). Letters A, B, C, D and V refer to different sampling dates on each year of the study period; see Table 3.

101x25mm (300 x 300 DPI)

Supporting Information, Text S1. Monte Carlo plug-in test.

When the pathogen compositions have an asymmetric relationship, e.g. N_2 was obtained by sampling in a population already sampled in the past, the sample in the past being N_1, an asymmetric alternative to the Chi-squared test can be applied, namely a Monte Carlo plug-in test. In this test, B samples $N_2^{(b)}$ ($b \in \{1,\ldots,B\}$, B large) are drawn under the multinomial distribution with size n_2 and with vector of probabilities $\hat{p}_1 = (1/n_1)N_1$, where $\hat{p}_1$ is the vector of variant proportions in sample 1. Then, the probabilities $m(N_2;n_2,\hat{p}_1)$ and $m(N_2^{(b)};n_2,\hat{p}_1)$ ($b \in \{1,\ldots,B\}$) are computed where $m(X;n_2,\hat{p}_1)$ is the probability that a vector drawn under a multinomial distribution with size n_2 and probabilities $\hat{p}_1$ is equal to X. Finally, the proportion of multinomial probabilities $m(N_2^{(b)};n_2,\hat{p}_1)$ ($b \in \{1,\ldots,B\}$) less than or equal to $m(N_2;n_2,\hat{p}_1)$ is the *p*-value of the test. If data are sparse, $\hat{p}_1$ is, under the null hypothesis an unbiased but relatively strongly varying estimate of p_1 and p_2 and, consequently, the Monte Carlo approximation of the distribution of the statistic $m(N_2;n_2,\hat{p}_1)$ is crude.

The failure of the Monte Carlo plug-in test was illustrated with the same simulation scheme than the one proposed to assess the performance of the Chi-squared test with sparse data (for the plug-in test, we used $B = 10^4$ samples).

Supporting Information, Text S2. Details about the analyzed datasets.

Magnaporthe oryzae. The dataset from Madagascar was collected in Andranomanelatra and is part of a pluri-annual population survey and was published in Saleh et al. (2014). The dataset from China was collected in Youle (Yunnan Province) and was published in Saleh et al. (2012). Fungal strains were obtained by monospore isolation from diseased samples as described by Silué and Nottéghem (1990) and stored as described by Valent et al. (1986).

Pseudomonas syringae. The Durance River valley is located in Southern France and drains the water collected over 14,250 km^2. The valley is composed of three areas characterized by a different hydrology, altitude, land occupation and climate; the upper part (UDR) in the French Alps, the middle part (MDR) and the lower part (LDR) downstream in the plains joining the Rhône River. The Durance River 302 km long essentially drains the flowing snowmelt from the mountains Alps in the UDR basin and the precipitation runoff along the valley in the smaller extend. The Durance River supplies the LDR area downstream where agricultural activity is intense and highly dependant of irrigation channels and water tables charged with alpine water sources. In this area, crops are mostly represented by arboriculture and horticulture surrounded by patches of deciduous plants, while lands in the UDR are mostly characterized by patches of open meadows and forests of Larches, Mountain and Arolla pines.

P. syringae samples were collected in rainwater at 10 different sites of LDR and UDR areas from 2007 to 2011. After sampling with sterile containers, samples were stored in a cooler for transportation and processed within the following day as described in Morris et al. (2008); a volume of water was concentrated by filtration. Processed samples were dilution-plated on KBC medium as described in Morris et al. (2008). After three days of incubation at room temperature, at least 30 strains per sample were purified on KB medium and test for absence of cytochrome c oxidase, fluorescence and arginine to determine the population size and between 10 and 15 strains were put in collection at –80°C (V/V 50% glycerol) for further genotypic characterization.

The *P. syringae* species complex is composed of several phylogroups (Berge et al., 2014). Morris et al. (2010) showed that phylogeny based on *cts* gene was reflecting that of the classification based on DNA-DNA hybridization studies. Within phylogroups, strains can be clustered in several clades. As described in Berge et al. (2014), two strains with a dissimilarity rate lower that 4.9% belong to the same phylogroups while they belong to the same clade if they are less than 2.0% dissimilar. The sequence of the *cts* gene of each strain was thus amplified as described previously from a pure suspension adjusted to 2×10^8 cells ml^{-1} with the primers described by Sarkar and Guttman (2004). PCR reactions were performed with a Qiagen Multiplex kit (Qiagen, Courtaboeuf, France) and their products were checked by electrophoresis in 2% agarose gels before sequencing.

Puccinia triticina. Wheat leaf rust was surveyed for seven consecutive years (2007-2013) in the region of Lomagne, southwestern France (Departements of Gers and Tarn-et-Garonne), a main regional area of winter wheat production; the usual crop rotation here is wheat-sunflower-wheat. The investigated zone (c. 350 km^2), centred on the Ancoupet farm (43° 57'N 0° 46'E, 116 m above sea level), is hilly (75 to 210 m above sea level). The landscape, without significant urbanized areas but small villages and isolated farms, is made of a mosaic of plots, often elongated rectangles, in average of 3 to 15 ha in area.
Every year, two "sentinel plots" (25 m x 50 m) of wheat cv. Galibier and Kalango, respectively, were delimited within bigger commercial plots located close to the Ancoupet farm (< 1 km). Cv. Galibier (registered in 1992 by Momont, Mons-en-Pévèle, France) has been the most grown cultivar in the investigated zone for, at least, the last decade. Cv. Kalango (registered in 2002 by Florimond Desprez, Cappelle-en-Pévèle, France) started to be grown in the investigated zone at the beginning of the experiment; however, the cultivar, not suited to the local conditions, was progressively dropped by farmers, so that in 2011 there was commercial plot left of this cultivar; it was kept in the experimental design, however, for the sake of continuity. Official rating (Geves / Arvalis-Institut du Végétal) for resistance to leaf rust is 2 for Galibier (very susceptible) and 3 for Kalango (susceptible). The two plots were submitted to the same cropping practices as the neighborhood, except they were left unsprayed with fungicides.

Wheat leaf rust was sampled in the two plots three to five times per year. During the cropping season, sampling dates depended on disease development. Thirty sampling points were marked with a stick according to a 5-m mesh regular grid (three lines of ten points each). At each sampling date and at each sampling point, a leaf bearing at least one sporulating lesion was excised and placed in an individual paper bag. At the first sampling date, minute lesions were checked when necessary using a glass lens to confirm they were actually caused by rust infection. During the intercrop season, wheat volunteers were surveyed short before sowing of the next

wheat crop. When volunteers with leaf rust lesions were present, a maximum of ten infected leaves were collected.

Field samples were processed according to standard techniques described in Goyeau *et al*., (2006, 2007) and Goyeau & Lannou (2011). In brief, a single-uredinium isolate was produced from each of the collected leaf. Pathotypes were determined by inoculating a differential set of wheat cultivars comprising 17 Thatcher differential lines with resistance genes to leaf rust *Lr1*, *Lr2a*, *Lr2b*, *Lr2c*, *Lr3a*, *Lr3bg*, *Lr3ka*, *Lr10*, *Lr13*, *Lr14a*, *Lr15*, *Lr16*, *Lr17*, *Lr20*, *Lr23*, *Lr26*, *Lr37*, the Australian cultivar Harrier carrying *Lr17b*, and the susceptible control Morocco. Infection types on the differentials were read 10 days after inoculation according to Stakman *et al.* (1962). An octal pathotype code (Gilmour, 1973) was assigned to each isolate.

Supplementary references:

Gilmour, J. 1973. Octal notation for designating physiologic races of plant pathogens. *Nature* 242.

Sarkar S. F., Guttman D. S. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal endemic plant pathogen. *Appl Env Microbiol* 70: 1999–2012.

Stakman E. C., Stewart D. M., Loegering W. Q. 1962. Identification of physiologic races of *Puccinia graminis* var. *tritici*. *U.S. Agric Res Serv* E617:1–53.

Supporting Information, Text S3. Incorporation of known binary virulences into the test

In some situations the underlying composition is filtered before being revealed by sampling. For instance, the two samples that are compared may have been obtained from two different host cultivars which have, each one, hampered the development of different variants. Another instance corresponds to the situation where each sample is formed by aggregating several samples obtained from several host cultivars with different resistances.

Such filtering induced by the cultivar has to be included in the analysis otherwise a significant difference between samples would be falsely detected (type I error). When the virulence of the variants (or the resistance of the cultivar) are known and binary, it is possible to include them in the test by considering the variants that are virulent for all the hosts. We provide a demonstration below.

Suppose that there are two groups of variants, say A and B, where variants of group A are virulent for two different cultivars, denoted by x and y, and variants of group B are not virulent for at least one of the cultivars. The two cultivars are assumed to be affected by the same inoculum composition described by the vector of probabilities $\pi = (\pi_A, \pi_B)$. Let $X = (X_A, X_B)$ and $Y = (Y_A, Y_B)$ be the two pathogen compositions sampled from the two cultivars. $X = (X_A, X_B)$ and $Y = (Y_A, Y_B)$ are independently drawn from multinomial distributions with vectors of probabilities proportional to $(\pi_A, \pi_B * v_x)$ and $(\pi_A, \pi_B * v_y)$, respectively, where * is the component-wise multiplication and v_x and v_y are the vectors of binary virulences of variants from group B over cultivars x and y, respectively. Then, given X_B and Y_B, the sub-compositions X_A and Y_A are independently drawn from multinomial distributions with the same vector of probabilities proportional to π_A. Therefore, X_A and Y_A can be compared with our test, or can be aggregated by a component-wise sum (i.e. $X_A + Y_A$) to form one of the pathogen compositions that are compared with our test.

Supporting Information, Text S4. Study of the selection of the weight w.

The weight w, which appears in the estimate $\hat{p}(w) = w\hat{p}_1 + (1-w)\hat{p}_2$ of p used in the GMCPIC test, was selected by minimizing the criterion provided by Equation (6) in the main text. This supporting information gives the distribution of the selected value $\tilde{w}$ of w in different simulation settings and shows that $\tilde{w}$ does not coincide in general with $\frac{n_1}{n_1+n_2}$ appearing in 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. Here, but also in all the computations that are presented in the article, w was optimized by (i) computing the criterion provided by Equation (6) for all values of w from 0.01 to 0.99 with a constant increment of 0.01 and (ii) selecting the value among these values leading to the minimum criterion value. In general, with B=10^4 and M=10^3, the criterion is a rather smooth function of the weight w; Hence, the minimization is rather stable with the simple minimization technique that was used. Our choice of B=10^4 and M=10^3 was governed by a trade-off between calibration accuracy and computation time. Obviously, the larger B and M, the more accurate the calibration. Based on our experience acquired from simulation studies, the values B=10^4 and M=10^3 lead to robust results in diverse situations with respect to the number of categories and the sample sizes. Thus, our advice is to use these values and, when only a few tests are made and computation time is not an issue, to increase the values of B and M to improve the test robustness.

We applied the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3, to simulation settings similar to those presented in Supporting Table S1, except that we varied the number of categories (i.e. the dimension of p_1 and p_2). Samples N_1 and N_2 were drawn from multinomial distributions with 3, 33 or 100 categories, varying sizes (n_1 and n_2 equal to either 10, 30 or 100), and randomly generated vectors of probabilities p_1 and p_2. For simulations with 3 categories, vectors p_1 and p_2 were obtained as follows: (i) the segment [0,1] was partitioned into 3 sub-segments by generating two variables U_1 and U_2 independently and uniformly distributed in [0,1] (suppose $0 \le U_1 \le U_2 \le 1$); (ii) then, we set $p_1 = (U_1, U_2 - U_1, 1 - U_2)$; (iii) finally, for the assessment of type I errors, we set $p_2 = p_1$, and for the assessment of powers, we set $p_2 = (V_{(1)}, V_{(2)} - V_{(1)}, 1 - V_{(2)})$ where $V_{(1)} = \min\{V_1, V_2\}$ and $V_{(2)} = \max\{V_1, V_2\}$, $V_1 = \min\{1, \max\{V_1', 0\}\}$, $V_2 = \min\{1, \max\{V_2', 0\}\}$, $V_1' \sim Normal(U_1, \sigma)$, $V_2' \sim Normal(U_2, \sigma)$, and $\sigma \sim Uniform([0.1,0.5])$. p_1 and p_2 were generated in the same way for simulations with 33 and 100 categories except that $p_1 = (U_1, U_2 - U_1, \frac{1-U_2}{31}, \dots, \frac{1-U_2}{31})$ and $p_2 = (V_{(1)}, V_{(2)} - V_{(1)}, \frac{1-V_{(2)}}{31}, \dots, \frac{1-V_{(2)}}{31})$ in the former case, and $p_1 = (U_1, U_2 - U_1, \frac{1-U_2}{98}, \dots, \frac{1-U_2}{98})$ and $p_2 = (V_{(1)}, V_{(2)} - V_{(1)}, \frac{1-V_{(2)}}{98}, \dots, \frac{1-V_{(2)}}{98})$ in the latter case.

Type I errors and powers were computed, for each number of categories and each pair of sample sizes (n_1, n_2), as the proportion of rejections over 400 repetitions. In these simulations, we considered cases with different sample sizes n_1 and n_2. Results are shown in Supporting Figure S3 and are consistent with observations made on the other simulation studies presented in the article.

The distributions of the optimal weight $\tilde{w}$ are shown in Supporting Figure S4 for each pair of sample sizes. We can see that the distribution of $\tilde{w}$ does depend on the pair of sample sizes, but does not significantly depend on whether the null hypothesis is true or not (simulations made for computing type I errors and powers led to similar distributions of $\tilde{w}$). We also clearly see that $\tilde{w}$ does not coincide in general with the weight $\frac{n_1}{n_1+n_2}$

appearing in 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. The weight $\widetilde{w}$ is generally larger than $\frac{n_1}{n_1+n_2}$.

Supporting Information, Table S1. Complementary simulation study. Type I errors and powers of the Monte Carlo Chi-squared test with B=10^4, the Fisher's exact test, its Monte Carlo version with B=10^4, and the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3. Here, Type I errors and powers were computed for samples N_1 and N_2 drawn from multinomial distributions with 3 categories, varying sizes (from 10 to 100), and randomly generated vectors of probabilities p_1 and p_2. Vectors p_1 and p_2 were obtained as follows: (i) the segment [0,1] was partitioned into 3 sub-segments by generating two variables U_1 and U_2 independently and uniformly distributed in [0,1] (suppose $0 \leq U_1 \leq U_2 \leq 1$); (ii) then, we set $p_1 = (U_1, U_2 - U_1, 1 - U_2)$; (iii) finally, for the assessment of type I errors, we set $p_2 = p_1$, and for the assessment of powers, we set $p_2 = (V_{(1)}, V_{(2)} - V_{(1)}, 1 - V_{(2)})$ where $V_{(1)} = \min\{V_1, V_2\}$ and $V_{(2)} = \max\{V_1, V_2\}$, $V_1 = \min\{1, \max\{V_1', 0\}\}$, $V_2 = \min\{1, \max\{V_2', 0\}\}$, $V_1' \sim Normal(U_1, \sigma)$, $V_2' \sim Normal(U_2, \sigma)$, and $\sigma \sim Uniform([0.1,0.5])$. Type I errors and powers were computed, for each sample size and each test, as the proportion of rejections over 2000 repetitions. Between brackets: p-value of the test of equality of the type I error to the value 0.05 (1st part of the table), or 95%-confidence interval of the power (2nd part of the table).

Criterion	Sample size	MC χ^2 test	Fisher test	MC Fisher test	GMCPIC test with stat. of Eq. (8)
Type I error	10	0.026 (<0.0001)	0.026 (<0.0001)	0.026 (<0.0001)	0.050 (0.96)
	20	0.039 (0.021)	0.039 (0.027)	0.039 (0.027)	0.050 (0.96)
	30	0.039 (0.021)	0.038 (0.011)	0.039 (0.021)	0.047 (0.50)
	50	0.041 (0.073)	0.044 (0.20)	0.044 (0.24)	0.048 (0.64)
	100	0.042 (0.090)	0.040 (0.045)	0.041 (0.058)	0.048 (0.64)
Power	10	0.23 (0.21,0.26)	0.23 (0.21,0.25)	0.23 (0.21,0.25)	0.27 (0.24,0.29)
	20	0.45 (0.42,0.47)	0.45 (0.42,0.47)	0.44 (0.41,0.47)	0.43 (0.41,0.46)
	30	0.59 (0.56,0.61)	0.59 (0.56,0.61)	0.58 (0.56,0.61)	0.58 (0.55,0.60)
	50	0.69 (0.66,0.71)	0.69 (0.66,0.71)	0.69 (0.66,0.71)	0.68 (0.66,0.71)
	100	0.85 (0.82,0.87)	0.85 (0.82,0.87)	0.85 (0.82,0.87)	0.83 (0.81,0.85)

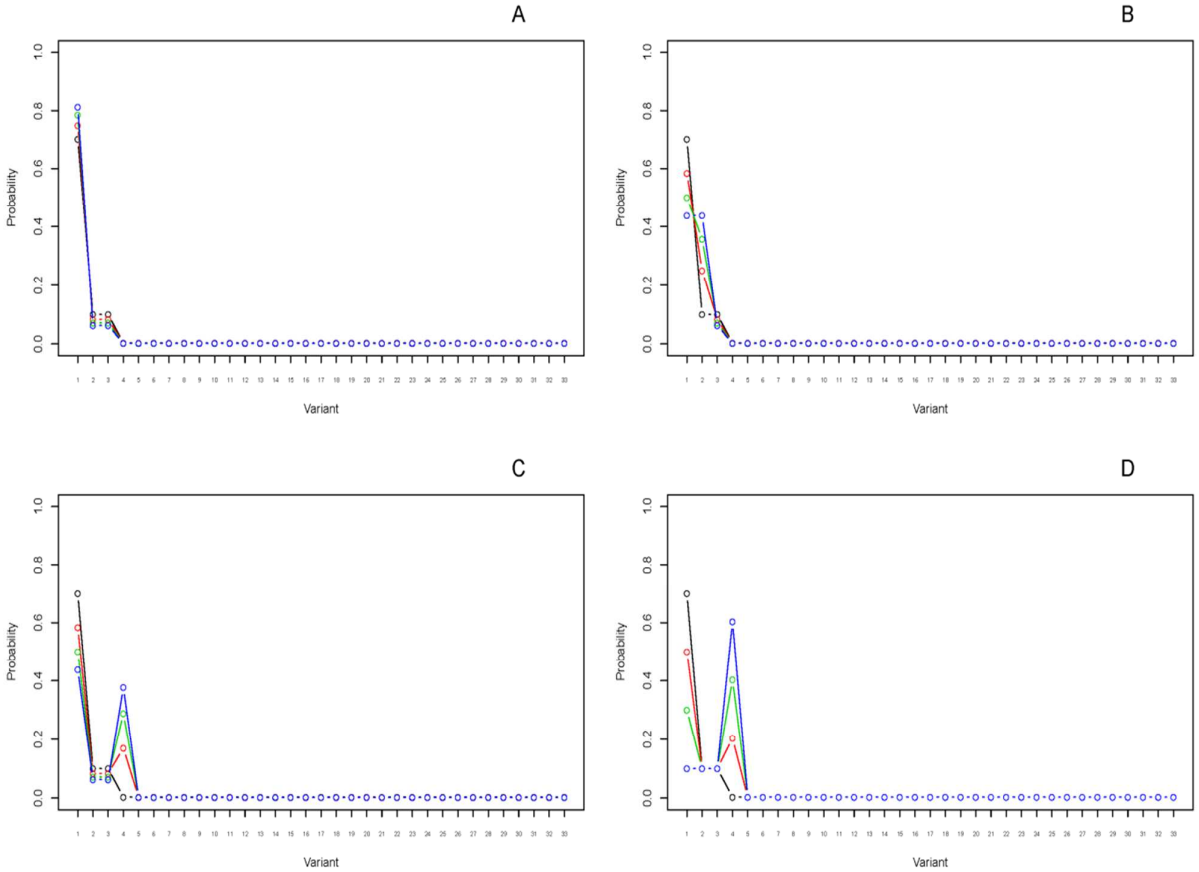
Supporting Information, Table S2. P-values obtained for the test calibrated at 0.05 and the multinomial density criterion. Letters A, B, C, D and V, in the row "Link tested", refer to different sampling dates on each year of the study period; see Table 3.

Year	2007			2008				2009			2010				2011			2012					2013	
Link tested	A-B	B-C	C-A	A-B	B-C	C-V	V-A	A-B	B-C	C-A	A-B	B-C	C-V	V-A	A-B	B-V	V-A	A-B	B-C	C-D	D-V	V-A	A-B	B-C
p-values																								
Galibier	.071	.002	NA	NA	.242	.242	.135	.573	.020	.058	.220	.331	.011	.011	.405	.429	.149	NA	NA	.016	.187	.030	.208	.275
Kalango	.048	.010	.207	.106	.285	.371	.189	.283	.037	.052	.167	.120	.092	.177	.031	.118	.180	.207	.011	.005	.266	.330	.301	.059
Galibier+Kalango	.102	.001	.012	.001	.236	.384	.196	.745	.026	.082	.206	.151	.003	.023	.528	.141	.062	.515	.059	.003	.300	.104	.574	.524

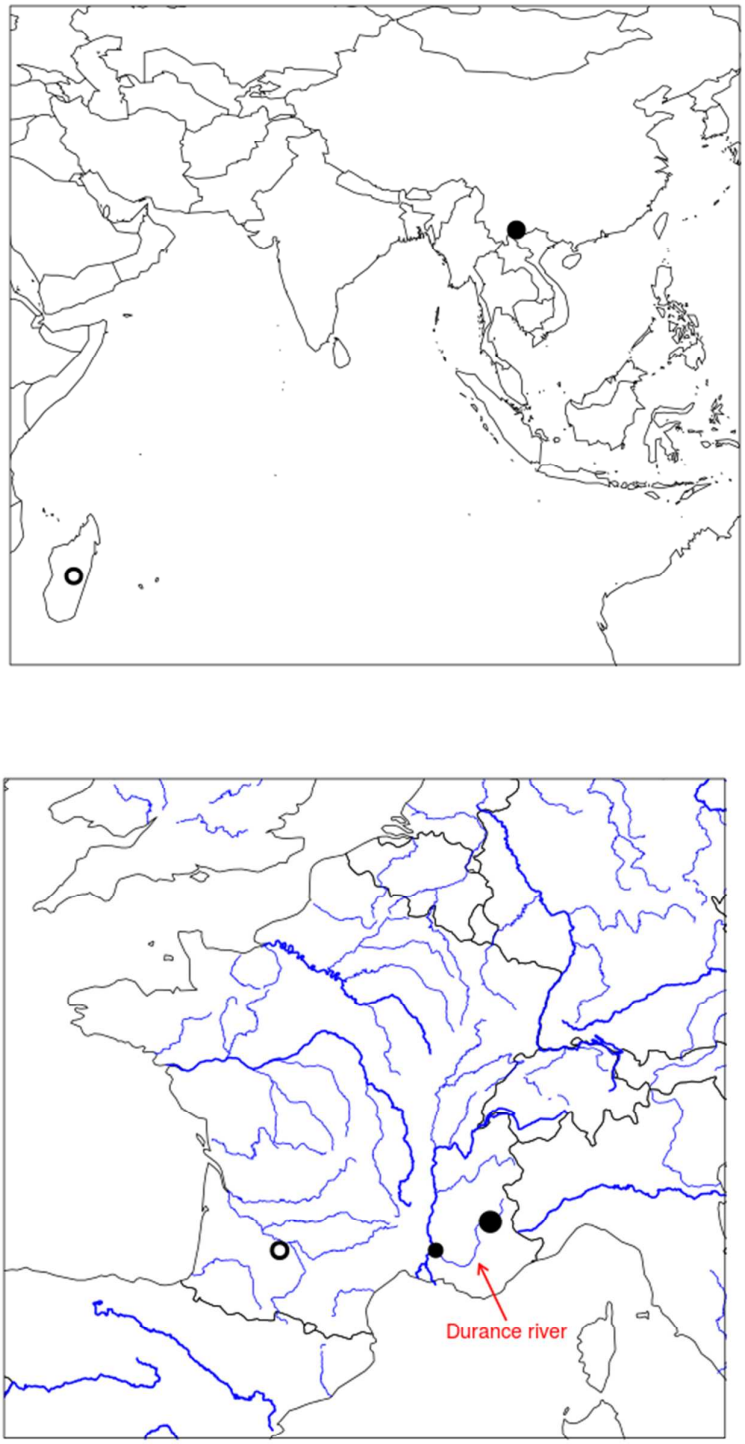
Supporting Information, Table S3. Pairwise comparison of the outcomes of the Monte Carlo Chi-squared test with B=10^4, the Fisher's exact test, its Monte Carlo version with B=10^4, and the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3, applied to *Puccinia triticina* data sets. The table below gives, for each pair of testing procedures, the number of times that the two procedures leads to inconsistent p-values, i.e. one p-value lower than or equal to the significance level 0.05 and the other p-value larger than 0.05. These numbers were computed by using the results of the tests applied to each pair of consecutive samples collected from the Galibier field only, from the Kalango field only, and from both the Galibier and Kalango fields (by discarding pathotypes that are not virulent for both Galibier and Kalango). There are 68 such pairs of consecutive samples. Thus, for instance, the Fisher test and the GMCPIC test lead to different conclusions in 10 cases over 68.

	Fisher test	MC Fisher test	GMCPIC test based on Eq. (8)
MC χ^2 test	3	4	9
Fisher test		1	10
MC Fisher test			9

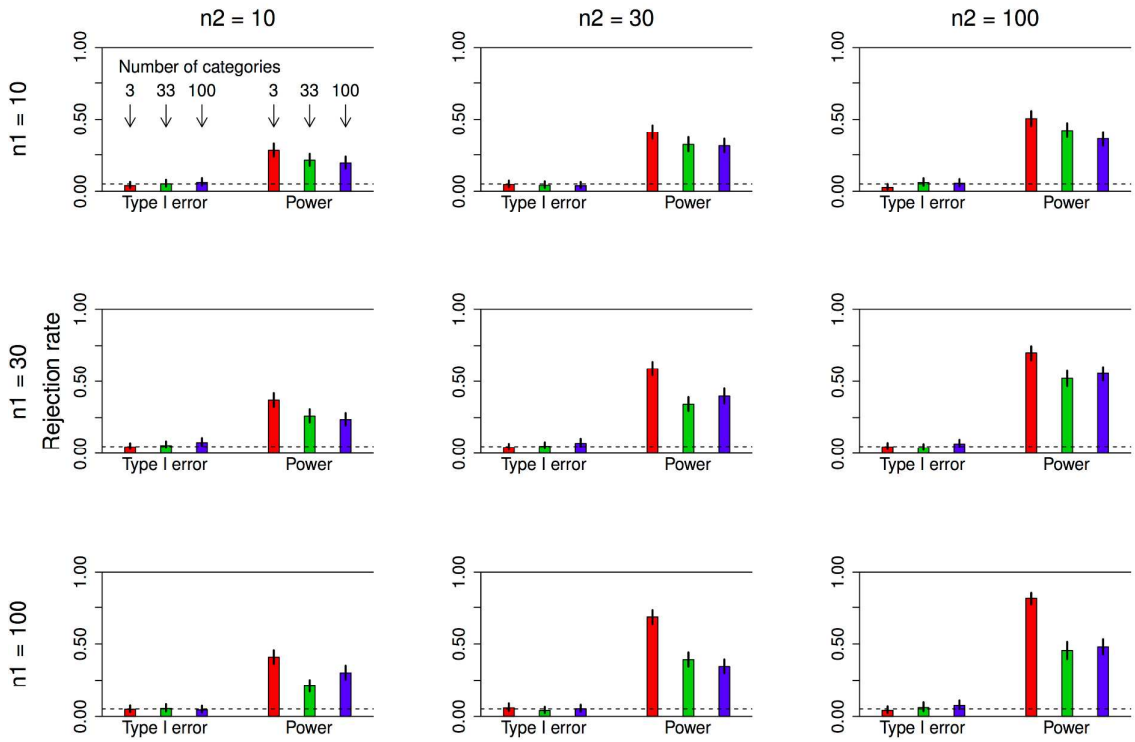
Supporting Information, Fig. S1. Vectors of probabilities used in the power analysis. The vector $p_1 =$ (0.70,0.10,0.10,0.10/30,...,0.10/30) is drawn in black in each panel. Values of $p_2 =$ (0.70+δ,0.10,0.10,0.10/30,...,0.10/30)/(1+ δ), (0.70,0.10+δ,0.10,0.10/30,...,0.10/30)/(1+ δ), (0.70,0.10,0.10,0.10/30+δ,...,0.10/30)/(1+ δ), and (0.70-δ,0.10,0.10,0.10/30+δ,...,0.10/30) are drawn in panels A, B, C and D, respectively, with δ=0.2 (red), 0.4 (green) and 0.6 (blue).



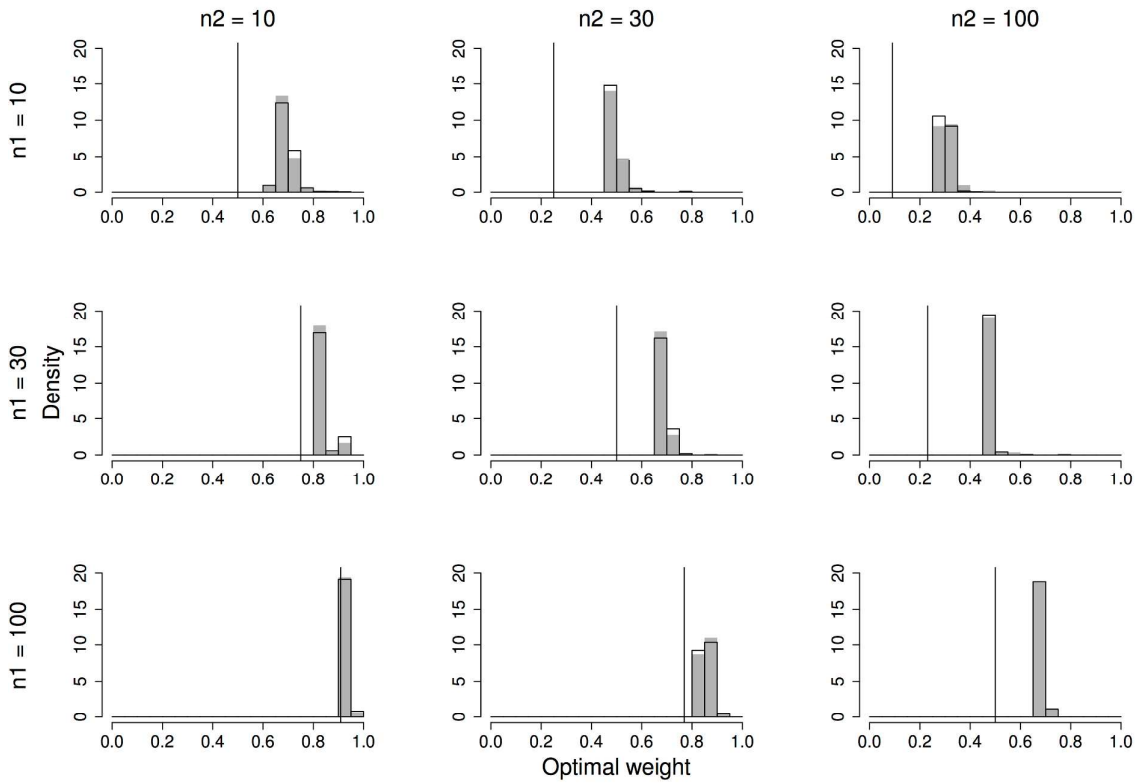
Supporting Information, Fig. S2. Locations of the sampling sites of *Magnaporthe oryzae* in Andranomanelatra (Madagascar; circle on the top panel) and in Youle (Yunnan Province of China; bullet on the top panel). Locations of the sampling sites of *Pseudomonas syringae* in the Upper Durance River basin (Southern Alps, France; large bullet on the bottom panel) and in Lower Durance River basin (at the junction of the Durance River and the Rhône River, France; small bullet on the bottom panel). Location of the sampling site of *Puccinia triticina* in Southwestern France (circle on the bottom panel). These maps were prepared with the R software (version 3.3.0, https://cran.r-project.org/).



Supporting Information, Fig. S3. Type I errors and powers of the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3, applied to the simulation settings described in Supporting Text S4. In these simulation settings, the sample size n_1 and n_2 were equal to 10, 30 or 100 and could be different, and the number of categories was equal to 3, 33 or 100. Each rejection rate was computed from 400 repetitions.



Supporting Information, Fig. S4. Distribution of the optimal weight $\widetilde{w}$ obtained in the application of the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3, to the simulation settings described in Supporting Text S4. In these simulation settings, the sample size n_1 and n_2 were equal to 10, 30 or 100 and could be different, and the number of categories was equal to 3, 33 or 100. The grey histograms correspond to the distribution of $\widetilde{w}$ when the two samples N_1 and N_2 were drawn with the same vectors of probabilities ($p_1 = p_2$; these simulations were used to compute test powers in Supporting Figure S3), whereas the black histograms correspond to the distribution of $\widetilde{w}$ when the two samples N_1 and N_2 were drawn with different vectors of probabilities ($p_1 \neq p_2$; these simulations were used to compute type I errors in Supporting Figure S3). Each histogram merges the distribution of $\widetilde{w}$ obtained from simulations with 3, 33 and 100 categories and was therefore obtained from 400×3=1200 repetitions. In each panel, the vertical line gives the value of the weight $\frac{n_1}{n_1+n_2}$ appearing in 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.



Supporting Information, Fig. S5. Distribution of the optimal weight $\widetilde{w}$ obtained in the application of the GMCPIC test with the statistic of Eq. (8), B=10^4 and M=10^3, to the *P. triticina* data set. This distribution was obtained by grouping the optimal weights $\widetilde{w}$ obtained from the application of the test to data from the Galibier field only, the Kalango field only and the Galibier and Kalango fields merged together.