Statistical analysis of the individual variability of 1D protein profiles as a tool in ecology: an application to parasitoid venom

Hugo Mathe-Hubert, Jean Luc Gatti, Dominique Colinet, M. Poirié, T. Malausa

Statistical analysis of the individual variability of 1D protein profiles as a tool in ecology: an application to parasitoid venom

H. MATHÉ-HUBERT,*†‡[1] J.-L. GATTI,*†‡[1] D. COLINET,*†‡[1] M. POIRIÉ*†‡[1,3] and T. MALAUSA*†‡[2,3]

*INRA, UMR 1355 Institut Sophia Agrobiotech, 06903 Sophia Antipolis, France, †Univ. Nice Sophia Antipolis, UMR 7254 Institut Sophia Agrobiotech, 06903 Sophia Antipolis, France, ‡CNRS, UMR 7254 Institut Sophia Agrobiotech, 06903 Sophia Antipolis, France

Abstract

Understanding the forces that shape eco-evolutionary patterns often requires linking phenotypes to genotypes, allowing characterization of these patterns at the molecular level. DNA-based markers are less informative in this aim compared to markers associated with gene expression and, more specifically, with protein quantities. The characterization of eco-evolutionary patterns also usually requires the analysis of large sample sizes to accurately estimate interindividual variability. However, the methods used to characterize and compare protein samples are generally expensive and time-consuming, which constrains the size of the produced data sets to few individuals. We present here a method that estimates the interindividual variability of protein quantities based on a global, semi-automatic analysis of 1D electrophoretic profiles, opening the way to rapid analysis and comparison of hundreds of individuals. The main original features of the method are the *in silico* normalization of sample protein quantities using pictures of electrophoresis gels at different staining levels, as well as a new method of analysis of electrophoretic profiles based on a median profile. We demonstrate that this method can accurately discriminate between species and between geographically distant or close populations, based on interindividual variation in venom protein profiles from three endoparasitoid wasps of two different genera (*Psyttalia concolor*, *Psyttalia lounsburyi* and *Leptopilina boulardi*). Finally, we discuss the experimental designs that would benefit from the use of this method.

Keywords: individual 1D SDS-PAGE, non-neutral markers, parasitoid venom, population proteomics, proteins, quantitative variation

Received 18 August 2014; revision received 12 February 2015; accepted 13 February 2015

Introduction

Deciphering the molecular basis of eco-evolutionary processes requires a range of informative molecular markers (Stapley *et al.* 2010; Davidson 2012). Common genetic markers include microsatellites, SNPs (single nucleotide polymorphisms) or markers obtained, for example, by the recent RADseq technique (Davey & Blaxter 2010). The large sets of markers identified are supposed to be randomly distributed in the genome, and most of them are expected to be neutral. However, neutral markers are not the most informative to measure the evolvability of phenotypic traits (Kirk & Freeland 2011; Karl *et al.* 2012). Moreover, it is not easy to link non-neutral SNPs to phenotypes, and an important part of the phenotypic variability is not determined by SNPs.

An important way to gather information on non-neutral molecular variation is to consider gene expression levels, as gene expression is a main step in the building of a phenotype (Diz *et al.* 2012) and a main source of intra- and interspecific phenotypic variability (Fay & Wittkopp 2008; Hodgins-Davis & Townsend 2009; Zheng *et al.* 2011). Variation in gene expression can be estimated through mRNA quantification using next-generation sequencing (NGS) approaches. However, mRNA quantification is only a tool to approximate the variation in protein quantities. Moreover, evolutionary signals are often clearer when considering protein rather than mRNA quantities (Feder & Walser 2005; Schrimpf *et al.* 2009; Guimaraes *et al.* 2014). Finally, natural selection

Correspondence: Hugo Mathé-Hubert, Fax: +33(0)4 92 38 64 01; E-mail: Hugomh@gmx.fr

[1]Evolution and Specificity of Multitrophic Interactions (ESIM)
[2]Biology of Introduced Populations (BPI)
[3]These authors codirected the work.

rather acts on proteins, more directly involved in the realization of a phenotype. Population proteomics has thus been developed as a relevant tool to measure non-neutral molecular variation occurring in the field (Biron *et al.* 2006; Karr 2008; Diz *et al.* 2012).

Proteins can be separated using different methods, from one (1D) or two dimensions (2D) SDS-PAGE, including 2D DIGE (Differential gel electrophoresis), to high performance chromatography. Their subsequent identification and quantification can be performed using methods based on mass spectrometry, such as proteomic shotgun combined or not with chemical labelling (Domon & Aebersold 2010; Slattery *et al.* 2012), that have been successfully used in an eco-evolutionary context (Burstin *et al.* 1994; López *et al.* 2001; Mosquera *et al.* 2003; Chevalier *et al.* 2004; López 2005; Diz & Skibinski 2007; Gonzalez *et al.* 2010; Rees *et al.* 2011; Papakostas *et al.* 2012; Slattery *et al.* 2012; Blein-Nicolas *et al.* 2013). However, these methods do not easily apply to population proteomics. Indeed, accurate estimation of interindividual variability relies on large sample sizes to ensure statistical accuracy (Crawford & Oleksiak 2007; Bolnick *et al.* 2011; Dall *et al.* 2012) whereas the number of individuals that can be analysed remains pretty low (mean number of 11.4 per group in the studies mentioned above). This leads to data sets composed of a large number of variables measured on a few individuals, that are thus statistically 'ill posed' and difficult to analyse. Altogether, proteomic techniques are too costly and time-consuming for being routinely used in large-scale population approaches. Besides, they require substantial protein quantities that cannot be obtained from small size individuals or individual tissues.

1D electrophoresis, a simple and low-cost method, has rarely been used to estimate interindividual variability in eco-evolutionary studies (see Bobkov & Lazareva 2012; Krishnan *et al.* 2012). This is probably because no automated method is available to rapidly and accurately analyse large numbers of individual gel lanes. Several image processing packages can correct for gel deformation, detect and align lanes, and transform each lane into an 'intensity profile' (the description of a lane through variations of intensity associated with bands). However, they reach their limit when protein samples are complex, mainly because of band overlapping. Indeed, automatic bands detection in each individual lane leads to recurring errors as soon as the intensities of the overlapping bands are variable. The matching of bands across the different lanes is often a second source of error. These errors call for manual corrections, incompatible with large sample size analyses.

We report here the development of a semi-automated method, implemented by a set of R functions, that allows analysis of individual 1D electrophoresis profiles obtained from digital analysis of gel pictures. This method is based on a semi-automated detection of 'reference bands' performed on a 'median profile', obtained from the whole set of individual lanes. Then, the intensities of these reference bands are recorded for each individual lane. Because it avoids the automatic detection of each band in each lane and thus the subsequent tedious manual screening of results, our method is more suitable for large sets of complex 1D profiles. The ultimate output of R functions mainly contains coordinates of reference bands as well as raw and normalized band intensities, as statistically analysable data sets. It is noteworthy that in addition to any protein sample analysable by 1D electrophoresis, the method might be used to compare high numbers of profiles of any kind, such as those obtained by HPLC, chromatography, ALFP or RFLP.

The method has been set up and tested on venom samples from three endoparasitoid wasp species, *Leptopilina boulardi* (Hymenoptera: Figitidae), *Psyttalia concolor* and *P. lounsburyi* (Hymenoptera: Braconidae), to assess its power for detecting intergroup structure and thus potential eco-evolutionary patterns. Endoparasitoid wasps are insects that lay eggs in the body of their host and develop at its expense, leading to its death. To ensure parasitism success, they largely rely on injection of venom inside the host along with the egg (Poirié *et al.* 2009), this venom being mainly composed of proteins. Interstrain and interindividual variability have been evidenced for venom of *Leptopilina* and *Psyttalia* endoparasitoids (Colinet *et al.* 2013). Moreover, the virulence of *Leptopilina* spp. was shown to evolve rather rapidly (Dupas *et al.* 2013), suggesting venom evolution might be involved in the observed variations of virulence. Using our method, we show that species, as well as geographically distant or close populations, can be discriminated based on individual venom profiles.

Material and methods

Samples and electrophoresis conditions

Origin and number of individuals, electrophoresis. Psyttalia lounsburyi populations from Burguret (Kenya) and Stellenbosch (South Africa) were collected in 2003 and 2005, respectively (Cheyppe-Buchmann *et al.* 2011), and reared since under laboratory conditions (Mathé-Hubert *et al.* 2013). *P. concolor* populations were collected in 2010 in Sicily and Crete. *Leptopilina boulardi* populations were sampled in 2010 in four sites of the Rhône Valley (France): Avignon, Eyguières, Sainte-Foy-lès-Lyon and Saint-Laurent-d'Agny. *P. lounsburyi* analyses were performed using reared females stored at −80 °C before analysis. *P. concolor* and *L. boulardi* analysed individuals

were females that were stored at −80 °C following field collection.

Sample preparation and analysis. *Leptopilina boulardi* venom reservoirs and *Psyttalia* spp. venom glands (*Psyttalia* reservoirs are mainly composed of muscle tissue) were dissected individually in 15 μL of insect Ringer solution supplemented with a protease inhibitors cocktail (PI; Roche). Residual tissues were removed by centrifugation (500 *g*, 5 min), and 10 μL of supernatant was mixed with an equivalent volume of Laemmli reducing buffer and heated (95 °C, 5 min). Proteins were separated by 1D SDS-PAGE electrophoresis using commercial gels (Any-kD Mini-PROTEAN® TGX™, Bio-Rad) for gel homogeneity. Silver stained gels (Morrissey 1981) were photographed (EOS-5D-MkII, Canon, Japan), and the high-resolution digital pictures (5626 × 3745 pixels; 16 bit; TIFF file) were analysed with the PHORETIX-1D software (TotalLab, UK). *Psyttalia* spp. venom samples were more variable in protein quantity than *L. boulardi* samples (data not shown). Therefore, for *Psyttalia*, three pictures were taken per gel at different times during the staining step and analysed with Phoretix 1D. The linear optical density (OD) range of the camera was around 2 (OD grey scale; T2115, Stouffer, IN, USA). Approximate cost of consumables was 2 € per sample.

Main steps of analysis of individual electrophoretic profiles

Adjusting for heterogeneity of gel migration and staining intensity. Step one—Alignment of individual protein profiles was performed with PHORETIX 1D. It relies on manual correction for relative mobility (hereafter abbreviated 'Rf'), using a set of chosen reference Rf positions (e.g. protein bands with few or no variation, easy to identify in all lanes) to deal with gel deformation such as 'gel smiling' (Biostep, 2008). Reference Rf positions are manually placed in all lanes of a gel picture and then linked to each other to create a Rf line. The software then calculates the Rf coordinates for all points between the Rf lines, thus allowing alignment of bands within and between gels (see Phoretix or GelAnalyzer information manuals). For *Psyttalia* spp., this step was performed for each of the three pictures per gel. As neither the gel nor the camera moved between pictures, an R script was used to map Rf lines from one analysed picture to the others (R script available in Appendix S1, Supporting information). Approximately 15 reference Rf positions were used in the analysis.

Step two—The local level of background intensity can vary between gels and lanes due to (i) variation in the loaded protein quantities, (ii) variation of intensity of adjacent bands or (iii) variations of gel staining and illumination. To assess whether the background introduces bias, we analysed each profile before and after removing the background by the 'rolling ball' method (based on the rolling of a 'ball' of 10 000 pixels of radius), following PHORETIX 1D instructions. The intensity profile of each lane (from each gel and picture) was thus exported two times, with and without the background, into a data table containing Rf positions and the intensity value for each pixel along the profile. Profiles with or without the background were, respectively, called 'B' and 'NB'.

Detecting and quantifying bands (corresponding to peaks on profiles). Step three—In this step, a median profile was 'calculated' and used to semi-automatically detect 'reference bands'. Then, the intensity of these bands was quantified in each individual lane using the height (maximum intensity of the band) and the volume (surface under the peak and between borders). Three profiles were obtained for each *Psyttalia* lane, corresponding to the three pictures. In this case, the method semi-automatically selects for each lane the profile showing the best staining level (i.e. compromise between detection of the weakest bands and absence of saturation of the most intense bands; see Fig. S1). This part of the analysis relies on the successive use of nine R functions (available in Appendix S1; see Figs 1 and S1). The function (*read.Profiles*) creates transformed profiles from the data table generated by PHORETIX 1D. Raw profiles are reduced to a set of points corresponding to the same Rf positions instead of pixel positions. These transformed profiles are used for the detection of reference bands (functions *Estim.2nd.derivative*, *Median.profiles_derivate*, *Detect.peaks*, *plot.profile* and *modify.peaks.manually*). The first two functions, (*Estim.2nd.derivative*) and (*Median.profiles_derivate*), calculate respectively for each Rf position (i) the second derivative of each profile and (ii) the weighted median of these second derivatives. To make the analysis more sensitive for the detection of rare bands, the median is weighted for each Rf position by the absolute value of the individual second derivatives. Thus, for each Rf position in the median profile, more weight is given to samples for which the second derivative has a signal, that is samples that display a band at the given Rf position. (*Detect.peaks*) function then detects local minima and maxima of the weighted median second derivative. Local minima reveal the position of bands common to individual profiles and they are used as reference peaks. Local maxima correspond to borders of these reference peaks. The function (*plot.profile*) plots notably the median second derivative calculated by the function (*Median.profiles_derivate*) and the position of detected peaks (Fig. S2, Supporting information). This helps to decide how reference peaks should be adjusted using the function

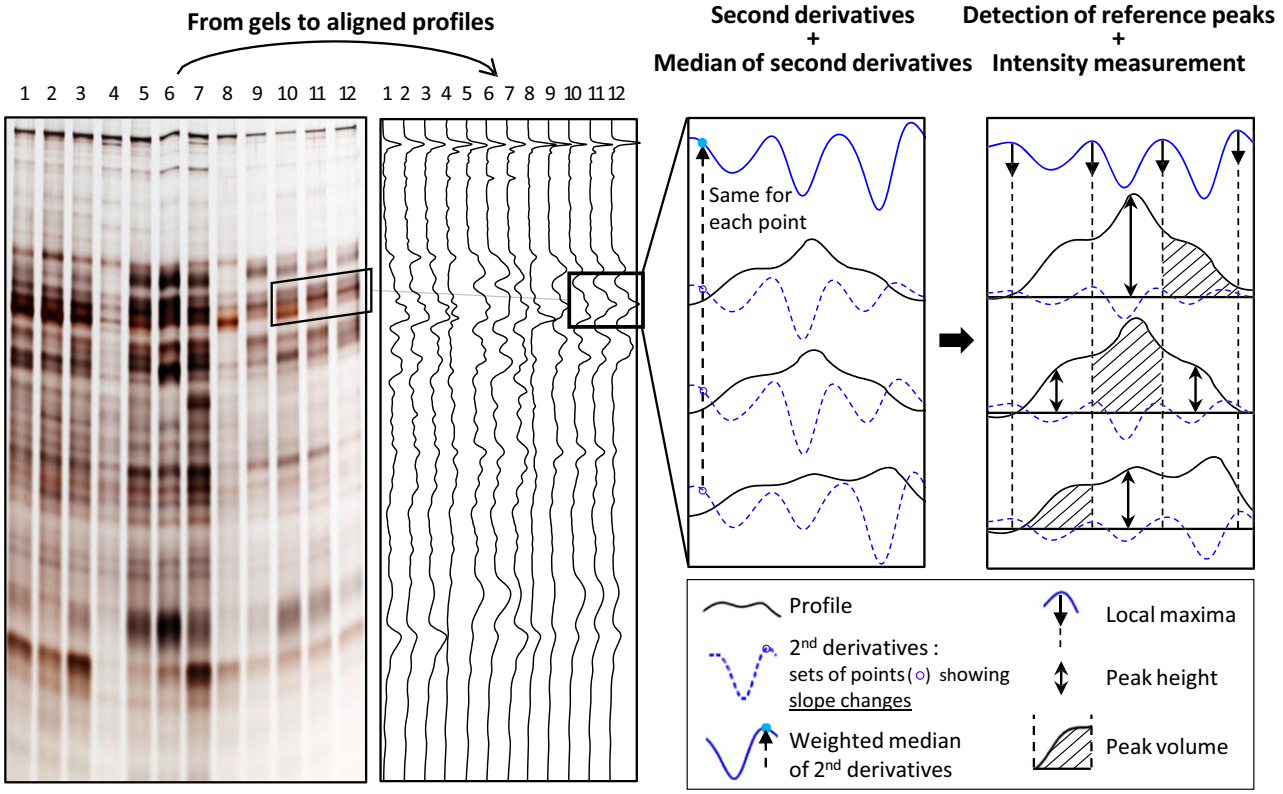


Fig. 1 Summary of the main analysis steps. The left part of the figure describes the preprocessing of gels pictures. The right part illustrates the analysis steps (details in the Material and Methods). Briefly, slope changes along the intensity profiles are calculated for each point of each profile, using second derivatives. The positions of the different points that correspond to the second derivatives are summarized by the blue dashed lines (only one point is represented per profile, as a circle). Second derivatives are then used to compute the weighted median (blue line) which summaries the slope changes of the analysed profiles. Local maxima of this weighted median of second derivatives correspond to the mean position of the inflection points of the profiles, and they are thus considered as borders of the 'reference bands' (vertical dashed lines with arrows at the top). These borders are used to measure the bands intensity, either with the height or the volume.

(*modify.peaks.manually*). Once borders are set for each reference peak of a profile, the seventh function (*Measure.peak*) records the coordinate of the top of peaks and their intensity (measure of the volume and the height of the peak). This procedure prevents two recurrent problems that occur in classical automated analyses of 1D electrophoresis. First, if two bands of unequal intensity partly overlap, the weaker band often appears as a shoulder of the stronger band instead of an independent peak. As band detection usually relies on the use of local maxima of the profiles (instead of the second derivative), only one large band is detected. Our method bypasses this problem by the use of second derivatives. Second, the automated step of band matching makes recurrent errors in case of polymorphism for the presence of bands that are close to each other. This problem is avoided by the use of fixed border coordinates, corresponding to the reference bands detected on the median profile, instead of the use of the band matching.

Taking into account the heterogeneity in the loaded protein quantity. Step four—This step aims at removing experimental variation to accurately analyse the variability in sample protein composition. The last R functions, (*select.photo*) and (*Compare.normalizations*), correct for the variability in the amount of loaded proteins, based on the analysis of different pictures of a gel (different stain levels). These functions can be used independently and (*select.photo*) is optional. The function (*select.photo*) selects for a given lane the picture that provides the best match between the intensity of the lane and the median intensity of all lanes (of all gels). This prevents heterogeneity in the level of saturation of lanes (see also Fig S1). The function (*Compare.normalizations*) normalizes the intensities using the 'limma' R package (Smyth 2005) to perform the three main normalization procedures (scale, quantiles and cyclic-loess; Smyth *et al.* 2003; Bolstad *et al.* 2003; Smyth 2005). The function 'removeBatchEffect' of the 'limma' package is also used at this step to remove

the gel effect, estimated through a linear model. Then, the function produces a graph comparing results obtained with different parameters combinations ('B' or 'NB' × peak height or volume × normalization by cyclicloess, quantiles or scale procedures).

Tests of the method accuracy

Sensitivity of the method. For the sensitivity analysis, a pool of venom was prepared from 13 venom glands of *P. lounsburyi* (see Sample preparation and analysis) and volumes equivalent to 0.1, 0.25, 0.5, 0.75, 1, 1.5 and 2 individual gland(s) were loaded twice on separate lanes of an SDS-PAGE gel. Following migration, the gel was silver stained and pictures were taken at three different staining levels (Fig. S3, Supporting information). This experiment was used to (i) analyse the saturation on profiles using different protein quantities and pictures, (ii) assess the power of the method to accurately describe variations of band intensity and (iii) check whether band intensity was equally sensitive to variations in protein quantity and staining duration. The pictures were analysed as described above without the functions *select.photo* and *Compare.normalizations* to obtain raw intensities.

To fulfil the first objective, the saturation range was characterized graphically (Fig. S3). For the second objective, a Box-Cox model (MASS package of the R 3.0.2 software; Venables & Ripley 2002) was fitted for each reference band and for the four parameters combinations ('B' or 'NB' × peak volume or height). Explanatory variables were (i) the loaded quantity, in equivalent of a venom gland, as a continuous variable, (ii) the picture as a discrete variable and (iii) the loaded quantity × picture interaction. To assess the power of the method, the coefficient of determination of each model was measured by the adjusted R^2, providing one coefficient of determination by band and parameters combination. This allowed comparison of accuracy of parameters combinations (Fig. S4, Supporting information). To compare the sensitivity of band intensity to staining and protein quantity, we used four coefficients per Box-Cox model to construct the two variables 'sensitivity to protein quantity' and 'sensitivity to staining duration'. The first coefficient of each model, the slope associated to the variation of protein quantity, represents the variable 'sensitivity to protein quantity'. In each model, three other coefficients describe the band intensity at the three staining duration (three pictures). The standard deviations of these three coefficients correspond to the variable 'sensitivity to staining duration'. We then checked graphically whether the relationship between the two constructed variables was linear, as expected if bands are equally sensitive to variation in protein quantity and staining duration.

Case studies: Test of the method. To test whether the method could discriminate between species and populations, we compared venom profiles between samples from different species (*P. concolor* and *P. lounsburyi*) and from different populations of a species (*P. concolor, P. lounsburyi* and *L. boulardi*). *Psyttalia* comparisons were based on analysis of 14 individuals for each of the two *P. lounsburyi* populations and the two *P. concolor* populations, loaded on four gels in a mixed design. *L. boulardi* sample sizes were 11, 20, 25 and 29 for 'Avignon', 'Eyguières', 'Sainte-Foy-lès-Lyon' and 'Saint-Laurent-d'Agny', respectively. Individuals were analysed on seven gels in a mixed design. *Psyttalia* spp. and *L. boulardi* gels were analysed separately.

Comparison of the accuracy of combinations of parameters—To compare the 'false' intersample variability (hereafter called noise) associated with the 12 parameter combinations ('B' or 'NB' × peak height or volume × the three normalization procedures), we calculated the ratio of intergroup variability to intragroup variability. This ratio is expected to decrease when the noise level increases, as noise introduces artificial interindividual variability but is not predicted to change interpopulation variability. The ratio was computed for each reference band, in the four studies designed to check intergroup variability [(i) *P. concolor* and *P. lounsburyi*, (ii) Cretan and Sicilian *P. concolor* populations, (iii) South African and Kenyan *P. lounsburyi* populations and (iv) the four *L. boulardi* populations] and a generalized linear model (GLM) with a gamma-distributed dependent variable was fitted to this ratio. Explanatory variables were the 12 parameters combinations and the identification number of the reference band. The gamma distribution was selected based on a Park test ('LDdiag' R package). Explanatory variables were tested with log-likelihood ratio tests. To compare the effect of each parameter combination, post hoc comparisons were performed with the 'multcomp' package (Hothorn *et al*. 2008).

Sensitivity of results to the combination of parameters—To check the power of the method to detect intergroup variation, we performed linear discriminant analyses (LDA) on normalized venom band intensities, for each case study and each combination of parameters. The sensitivity of LDA to the combination of parameters was tested for each case by selecting and comparing two LDAs. The first LDA used the combination of parameters with the highest ratio of intergroup versus intragroup variability. The second LDA used the combination of parameters that had both (i) a high inter- versus intragroup ratio and (ii) one of the most different combination of parameters compared to the first LDA. The parameters combinations

Table 1 Combinations of parameters used for the LDAs in the four case studies. The parameter combination used for each of the two LDAs of each case study is provided in the three last columns. The column 'Background' indicates whether the background was subtracted (NB) or not (B). The column 'Intensity' indicates whether band intensity was measured with the peak height (H) or the peak volume (V)

		Parameters combination		
Case studies (Groups)	LDA	Background	Normalization	Intensity
Psyttalia spp.	1	NB	Quantiles	H
(*P. lounsburyi*, *P. concolor*) Fig. 4a	2	NB	Scale	V
P. lounsburyi	1	NB	Scale	H
(South Africa, Kenya) Fig. 4b	2	NB	Quantiles	V
P. concolor	1	NB	Quantiles	H
(Crete, Sicily) Fig. 4c	2	B	Cyclic-loess	V
L. boulardi	1	B	Scale	H
(Avignon, Eyguières, Ste Foy Lès Lyon, St Laurent d'Agny) Fig. 5	2	NB	Quantiles	V

of the two LDAs for each case study are summarized in Table 1. In each case, the two LDAs were compared using correlations of bands to LDA axes, which allow identifying bands that show intergroup variation. More precisely, we checked whether each band was similarly correlated to axes of the two LDA. The correlations of bands to the LDA axis were tested by a Spearman rank correlation test with a Bonferroni correction for the number of reference bands (38 for *Psyttalia* spp. and 32 for *L. boulardi*). LDAs were performed and tested with the ADE4 R package (Dray & Dufour 2007).

Results

Sensitivity and validity of the method

Comparison of electrophoretic profiles between individuals requires (i) the accurate alignment of profiles and (ii) the detection of variation in staining intensity as a reliable evaluation of the protein amount per band. The first requirement was fulfilled by the profile analysis performed using Phoretix 1D and the following treatment of data based on R functions, combined with a manual adjustment for some of the reference bands based on the constructed 'median profile'. For the second requirement, a sensitivity analysis was performed using protein profiles corresponding to different amounts of the same sample and three pictures corresponding to different stain levels. Although a low saturation of the band intensity occurred in the tested range of protein amounts (Fig. S3), it did not prevent detection of variations in intensity (Fig. S4). Indeed, the adjusted R^2 that describe for each band the amount of variation in intensity explained by the loaded quantity and the staining duration were overall much higher than 0.95 before background subtraction and ranged between 0.9 and 1 after background subtraction. This indicates that background subtraction marginally reduced the sensitivity (Fig. S4).

We also used this experiment to check whether the shooting of gels at different staining times was a useful approach to deal with the variation in loaded protein amounts, thus allowing comparing lanes of similar stain levels. This required that the band intensity varied similarly with changes in the staining duration and the loaded protein amount. A strong linear relationship was indeed observed between the variables 'sensitivity to protein quantity' and 'sensitivity to staining duration', although the correlation was lower after background subtraction (Fig. 2).

Comparison of the accuracy of combinations of parameters on real data sets

In this section, we have used real data sets of venom of three wasp species to compare the 12 combinations of parameters with respect to their capacity to identify interpopulation differences.

The developed analysis was used on two independent data sets. Venom profiles of *P. lounsburyi* and *P. concolor* were analysed together, providing 38 reference bands, while analysis of *L. boulardi* profiles led to identification of 32 reference bands. To compare the noise level associated with different parameter combinations, the ratio between the inter- and intragroup variability was computed for each parameter combination and each band, using data from the four cases (venom profiles comparison between *Psyttalia* species and between populations of *P. concolor*, *P. lounsburyi* and *L. boulardi*; see *3b.* in material and methods). A significant effect of the two explanatory variables ('combination of parameters' and 'reference band') was observed in all cases (Table S1, Supporting information), with post hoc comparisons evi-

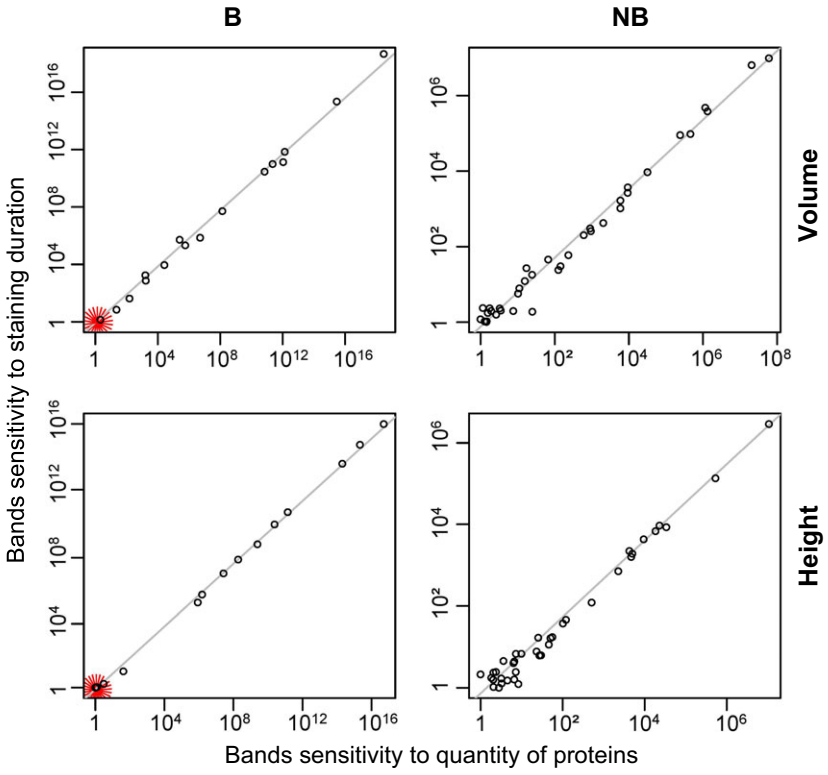


Fig. 2 Linearity of the effect of the protein quantity and of the staining duration on the band intensity. Four parameters combinations [with (B), or without (NB) background; detection of height (H) or volume (V)] were tested. For each reference band, a model was fitted to explain the band intensity by the protein quantity and the picture identifying number. Each point represents one reference band. The *x*-axis and *y*-axis correspond to the variable 'sensitivity to the protein quantity' and 'sensitivity to the staining duration', respectively. For logarithmic scale, the two variables were transformed by subtracting each value by the minimum of the variable and adding one. The red spiny shapes at the lower left of the two left graphs represent overlapping points (one 'spine' per overlapping point).

dencing differences between specific parameters combinations (Fig. 3). Although the best parameter combination was not the same in all cases, it always involved the 'peak height' quantification. Moreover, similar trends occurred for (i) *P. lounsburyi* and *P. concolor* and (ii) *P. lounsburyi* populations (Fig. 3a and b), with higher ratio values without background subtraction and using a quantiles normalization. However, striking differences were also observed. For example, although the combination 'No background' ('NB'), height, with quantiles normalization was among the best ones in all situations involving *Psyttalia* spp., it was one of the worst when considering *L. boulardi* data (Fig. 3). This absence of common trend is probably due to interactions between the combination of parameters and the characteristics of the bands showing a high intergroup variability (e.g. size, shape flat or pointed and distances to adjacent bands).

Background subtraction was shown to introduce some noise in the sensitivity test (Fig. S4), performed with a unique venom sample, while it seemed to decrease noise when using data from *Psyttalia* species (Fig. 3). This suggests that local background subtraction may improve analyses in some cases. Although results seem to be rather robust to the combination of parameters, it is then advisable to compare results obtained with different combinations of parameters as far as new samples/biological models are to be analysed.

Nonetheless, the ratio values were clearly higher for comparisons of (i) *P. lounsburyi* and *P. concolor* and (ii) *P. lounsburyi* populations whatever the combination of parameters, suggesting a higher intergroup differentiation. Similarly, the lowest ratio values were observed for comparison of *P. concolor* populations whatever the combination of parameters.

Case studies: Test of the method

Linear discriminant analyses (LDA) were used in the four comparisons to test the power of the method to detect intergroup variation. All combinations of parameters allowed the detection of highly significant ($P < 10^{-3}$) intergroup differences in venom in the four cases. Pattern detection was thus not hampered by the noise due to a combination of parameters despite the small sample sizes (11–20 individuals per population).

To compare LDA layouts obtained with different parameters (Table 1), we used the correlations of bands with LDA axes, which allow identifying bands that show intergroup variation. When venom composition was compared at the species level (*P. concolor* vs. *P. lounsburyi*), 18 bands were significantly correlated with the LDA axes and two bands with the axis of the first LDA only (Fig. 4a).

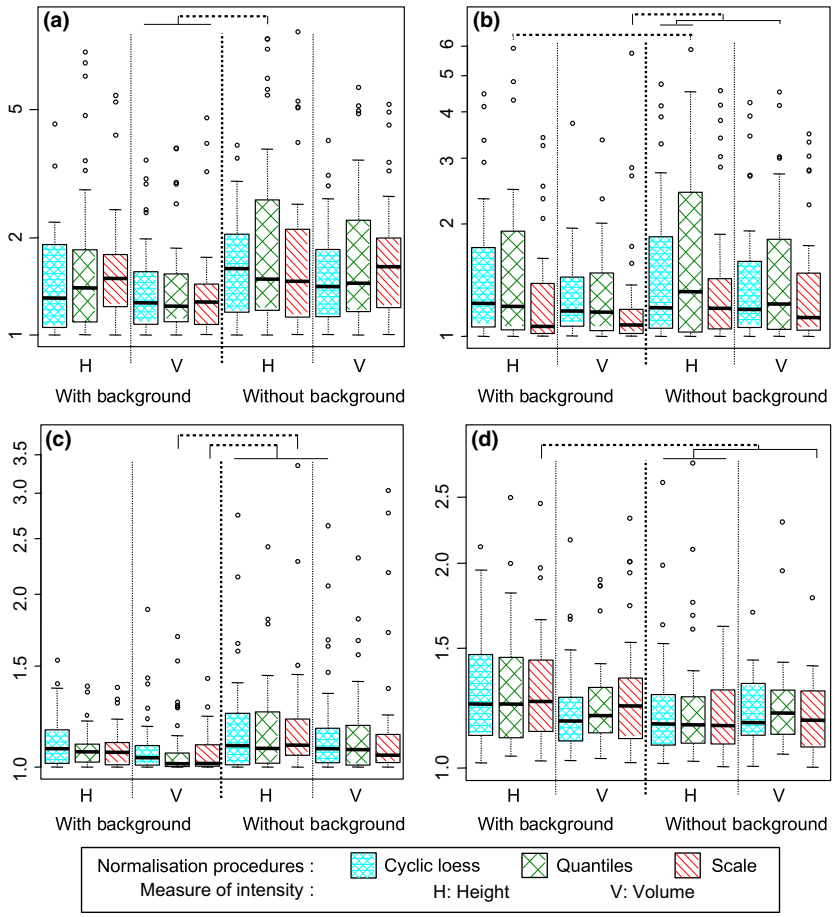


Fig. 3 Comparisons of the accuracy of combinations of parameters. The ratio (Intergroup variation/Intragroup variation) +1 is shown for each band and parameters combination. Boxplots compare the ability of parameters combinations to detect venom-based intergroup structure between (a) *P. lounsburyi* and *P. concolor*, (b) *P. lounsburyi* populations, (c) *P. concolor* populations and (d) *L. boulardi* populations. Horizontal lines above boxplots indicate significant differences. Thin lines group combinations that are related by a dotted thick line indicate significant differences between parameters combinations on the left *versus* the right side of this dotted thick line. This graph resembles graphs produced by the R function (*Compare.normalizations*), except for the statistical significance part.

At the population level, for *P. lounsburyi* (South Africa vs. Kenya), four bands were significantly correlated with the discriminant axis of the two LDA, while six bands were correlated with the axis of the first LDA only (Fig. 4b). For *P. concolor* (Cretan vs. Sicilian populations), only the band #6 was significantly correlated with axis of both LDAs, one band being correlated with the axis of the second LDA only (Fig. 4c). This low number of discriminating bands is in agreement with the low intergroup versus intragroup variation for *P. concolor* populations (see above). Interestingly, variation of intensity of band #6 was detectable (Fig. 4d) although this band is among the weakest ones in terms of intensity (Fig. S5a, Supporting information). This indicates that the method is powerful enough to detect intergroup differences based on a low number of bands of low intensity, even using small sample sizes.

In the comparison of venom proteins of *L. boulardi* French populations (Avignon, Eyguières, Sainte-Foy-lès-Lyon and Saint-Laurent-d'Agny), the first and second discriminant axes were inverted, probably because eigenvalues of first and second axes were similar (0.78 and 0.72 in the first LDA, 0.79 and 0.70 in the second LDA). Once the two axes of the second LDA were inverted, the global patterns of LDAs 1 and 2 were qualitatively similar, the main difference being a slight clockwise rotation from LDA 1 to LDA 2 (Fig. 5). Seven bands were significantly correlated to the first axis of LDA1 and the second axis of LDA2 (horizontal axes; arrows with triangle or star at the origin and at the end, Fig. 5b), while 12 bands were correlated to one of these two axes only (arrows with triangle or star at one of the two extremities only; Fig. 5b). For the vertical axes, one band was correlated to the second axis of LDA1 only (arrow with a cross at the origin only, Fig. 5b) and two bands to the first axis of LDA2 only (arrows with a cross or a star at the end only, Fig. 5b). This lack of conservation of the significance level is mainly due to the slight clockwise rotation from LDA 1 to LDA 2.

Overall, variability in band correlations to discriminant axes was mainly found for bands that were poorly or not correlated to LDA axes (Figs 4 and 5).

In summary, the structures between groups described by the LDAs are well conserved through the different parameters combinations, meaning that the choice of a combination only marginally affects the results, provided that the selected combination is one of those displaying the highest 'intergroup vs. intragroup variation'.

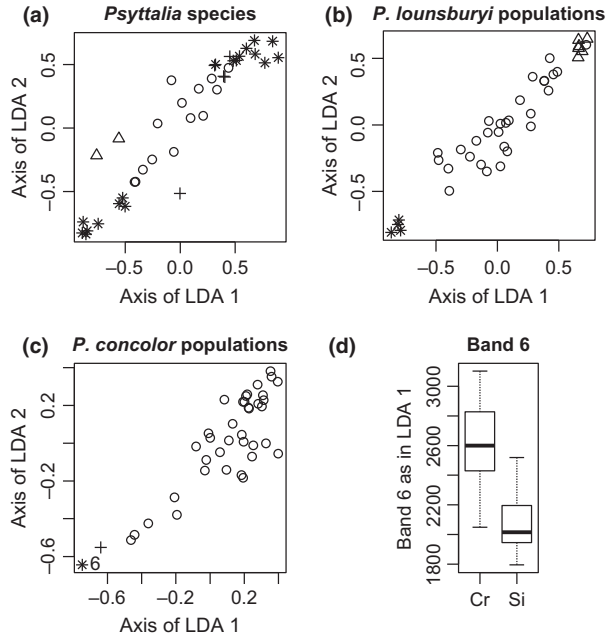


Fig. 4 Discriminant analyses on *Psyttalia* spp. venom. The graphs compare results of LDAs performed on data from *Psyttalia* spp. with different parameter combinations. (a) Comparison between *P. lounsburyi* and *P. concolor*, (b) Comparison between South African and Kenyan *P. lounsburyi* populations, (c) Comparison between Cretan (Cr) and Sicilian (Si) *P. concolor* populations. There is only one discriminant axis per LDA as only two groups are to be discriminated. Plots describe the adequacy between the first and second LDA obtained with different combinations of parameters (see Table 1): *x*- and *y*-axes correspond to the correlation between the LDA axes and the band intensities obtained with the combination of parameters used in the LDA. Symbols 'stars', 'triangles', 'crosses' and 'circles' correspond, respectively, to bands significantly correlated to (i) the axis of the two LDAs, (ii) the LDA1 axis, (iii) the LDA2 axis, (iv) none of the axis. (d) Intensity of band #6 (only band significantly correlated to both LDA axes) in Cretan and Sicilian *P. concolor* populations measured with the same combination of parameters as for the LDA1 (Table 1). The position of band #6 is indicated on Fig. S4a by the arrows.

Discussion

The method

The last 10 years have seen an increasing number of applications of proteomics to a wide range of biological fields such as behavioural ecology, molecular ecology or evolution (Navas & Albar 2004; Biron *et al.* 2006; Karr 2008; Melzer *et al.* 2008; Rees *et al.* 2011; Diz *et al.* 2012; Valcu & Kempenaers 2014). Proteomes show large differences in protein abundance which carry biological information not accessible through genomics or transcriptomics (Feder & Walser 2005; Maier *et al.* 2009; Laurent *et al.* 2010; Diz *et al.* 2012). Here, we have developed and tested a method allowing the semi-automated analysis of 1D protein electrophoresis profiles as an helpful tool to address eco-evolutionary questions.

Because this method relies on simple 1D electrophoresis, it makes it possible and affordable to analyse the large numbers of individuals required for accurate estimation of interindividual variation, a keystone to ecological and evolutionary studies (Crawford & Oleksiak 2007; Bolnick *et al.* 2011; Violle *et al.* 2012). For instance, although we present here significant results from analysis of a low number of individuals and populations, gathering information from a higher number of locations and individuals will be required to identify the causes of the observed population structure. Results are currently being obtained from experimental evolution experiments using the developed method, which already required analysis of the venom content of more than 1000 parasitoid individuals in a short period of time.

A critical step in biochemical experiments is to normalize sample quantities to make them comparable. This step is tedious and can require a large amount of each protein sample, as for parasitoid venom. Here, we show that sample quantities can be homogenized *in silico* by analysing gels at different staining level. The number of pictures required per gel depending on the variability in protein quantities (Appendix S2, Supporting information). This makes it possible to compare lanes of similar stain level, equally prone to saturation, following a last normalization step to deal with the remaining variability.

Another important feature of the method is that it bypasses the problem of recurrent errors arising from automated band detection on gel lanes, by detecting reference bands on a median profile and by recording the intensity between the borders of these reference bands. However, one constraint is that the method accuracy depends on the precision of the detection of the reference bands which itself depends on the homogeneity of the profiles. For example, it could be preferable to analyse the intraspecific variability of *P. lounsburyi* and *P. concolor* venom profiles separately (with one set of reference bands per species) because the venom profiles of these species largely differ.

Analysis of 1D electrophoresis patterns has some intrinsic limitations, such as the absence of detection of low-abundant proteins, or the problem of band overlapping with adjacent bands that can result in artefactual correlations between bands intensities. Notably, statistical approaches to handle multicollinearity and its possible modulation by the choice of the combination of parameters are discussed in Appendix S2. Another limitation is that a large part of the post-translational modifications is not detected by 1D electrophoresis-based

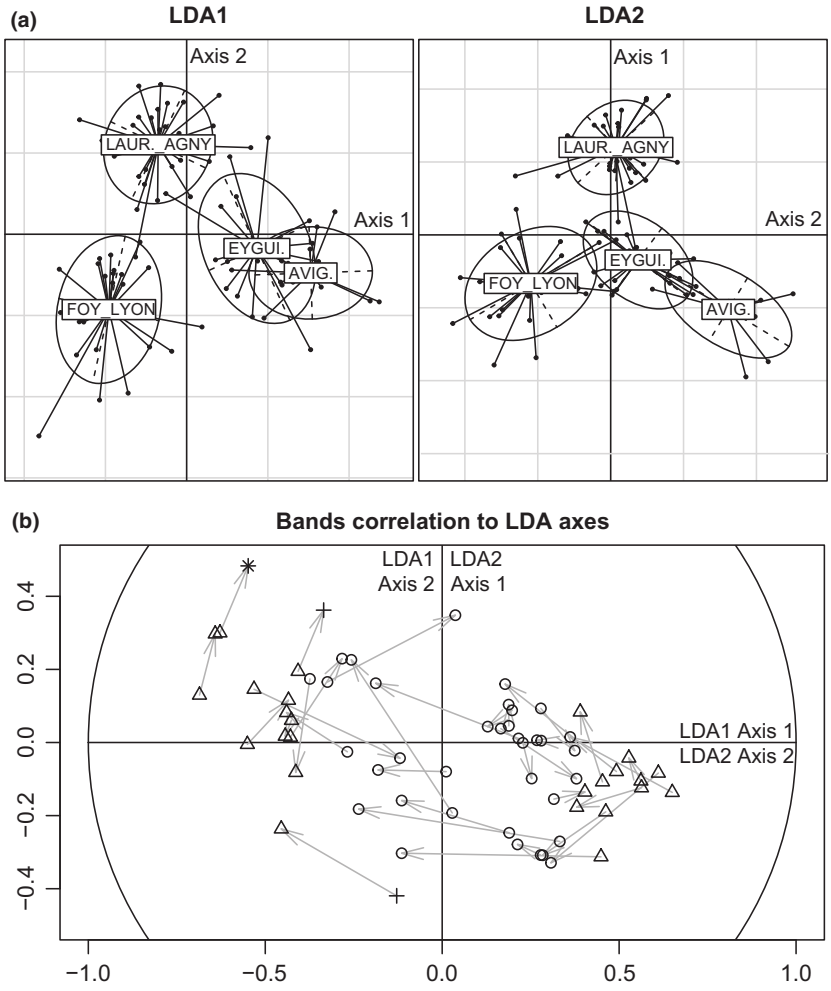


Fig. 5 Discriminant analyses on the venom of the four populations of *L. boulardi* venom. (a) Results of the first and second LDAs performed on the four *L. boulardi* populations using different parameter combinations (each point show the position of one individual on discriminant axes). To make the two LDAs comparable, LDA2 axes have been transposed. (b) Correlation of bands to axes of the two LDAs. As in Fig. 4, coordinates on axes indicate the correlation coefficient. The origin and end of the arrows correspond to the correlations of the bands with axes of the LDA1 and of the LDA2, respectively (with transposed axes). Thus, symbols at the origin and end of arrows describe the significance of band correlation to axes of the LDA1 the LDA2, respectively. Symbols 'stars', 'triangles', 'crosses' and 'circles' correspond to bands significantly correlated to (i) horizontal and vertical axes, (ii) horizontal axis only, (iii) vertical axis only, (iv) none of the axes, respectively. Changes induced by differences in the parameter combination are evidenced by the length of arrows and the concordance of symbols at the origin and the end of arrows. Most arrows indicate a clockwise rotation.

studies (Sickmann *et al*. 2001; Jensen 2006; Jacob & Turck 2008; Karve & Cheema 2011). Future technological advances will hopefully overcome part of these limitations. Overall, the best statistical approaches for analysis of the data sets obtained with the method seem to be multivariate analyses. Alternative approaches, R packages and softwares are also discussed in Appendix S2.

Although the individual variation in protein expression cannot be entirely described by 1D electrophoresis, our method still identified enough variation to perform accurate analyses of the populations structure. This opens the way to a rapid and cost-effective analysis of natural variability of protein expression.

The example of parasitoid venom analysis

Endoparasitoid venom was a good protein sample to test the developed method because it is a protein-composed fluid of medium complexity, easy to collect through individual dissection. Our results demonstrate that the method can accurately discriminate between species, as well as geographically distant or close populations of a species, through estimation of interindividual variation in protein quantities. They also provide the first evidence that endoparasitoid populations can be discriminated on the basis on their venom composition.

At the intraspecific level, laboratory populations of *P. lounsburyi* were the most differentiated, followed by field populations of *L. boulardi* and *P. concolor*. Although differences in *P. lounsburyi* may also be attributed to genetic drift and/or rearing effects, they are in agreement with the strong neutral differentiation in the field between Kenyan and South African populations (F_{ST} = 0.4; Cheyppe-Buchmann *et al*. 2011). In contrast, Mediterranean *P. concolor* populations have a low level of genetic divergence (Karam *et al*. 2008). Finally, the four *L. boulardi* populations could all be discriminated by their venom content although some of them were sampled in close locations.

Explaining an observed quantitative variation in a protein band detected on 1D SDS-PAGE is not straightforward. Indeed, variation can be due to post-transla-

tional differences (frequency of isoforms) or to genetic or epigenetic differences in the regulation of protein expression. Protein profiles can also be influenced by individual variables (e.g. sex, age) and environmental conditions. Their geographic variation thus probably results from both plasticity and genetically based modifications that may be neutral or involve local adaptation. As we only aimed here at measuring and statistically analyse the naturally occurring proteomic variability, the presented data do not discriminate between neutral genetic variation and local adaptation as main factors of the venom-based population structure. Disentangling the causes of the observed variability will require specifically designed approaches as for instance common garden experiments or/and the coupling of protein profiles characterization with neutral markers. Once protein bands associated with population divergence have been identified, such as the band #6 of *P. concolor*, they can be more thoroughly investigated to identify the molecular bases of the observed variability. As one protein band usually contains several proteins, complementary approaches to the global analysis may be useful, for example the use of tools specific to already characterized proteins or the knock-down of specific genes through RNA interference (e.g. Li *et al.* 2012; Colinet *et al.* 2014). The primary identification of protein markers may thus lead to the final characterization of proteins involved in population structuring or associated to a specific phenotype.

Perspectives: Application for ecology and population studies

Not all protein samples may be suitable for analysis using the developed method. Indeed, it relies on the assumption that abundant proteins, easily detected in 1D SDS-PAGE, are those that display eco-evolutionary patterns of interest. Specific tissues, such as glands or fluid compartments, have often been successfully analysed by classical 1D electrophoresis [see for instance: venom (Colinet *et al.* 2013), vitelline envelope (Aagaard *et al.* 2006), seeds storage proteins (Bobkov & Lazareva 2012), eyes tear film (Green-Church *et al.* 2008), liver and brain extracts (Gonzalez *et al.* 2010; Fig. S2)], but the interest of this approach remains to be tested for other tissues or small entire organisms. The use of individual 1D electrophoresis in an ecological or evolutionary context has only been reported, to our knowledge, for snake venom (Angulo *et al.* 2007; Alape-giron *et al.* 2008; Sanz *et al.* 2008; Núñez *et al.* 2009) and rodent seminal fluid (Ramm *et al.* 2009).

The method generates a set of markers independently from prior knowledge on the studied species. It is thus suitable for ecological studies on nonmodel organisms. Moreover, the protein content of chosen bands can be easily characterized even for nonmodel organisms (Armengaud *et al.* 2014), paving the way for identification of sets of proteins and genes involved in an adaptive trait. This method can thus be useful in the process of identifying environmental or geographic factors that explain variation in protein expression, or of testing hypotheses based, for instance, on experimental evolution studies.

Finally, such markers of protein expression may be helpful to track and predict the effects of global change on population dynamics by providing data on adaptive and plastic responses to environmental conditions (Reusch & Wood 2007). They could also be valuable in the context of trait-based community ecology that aims to explain the structure of community of species by their traits. For instance, the data produced could be integrated in the framework that relies on estimation of traits variability at each organizational level (from individuals to communities of species) to determine which level mostly impacts the trait in the community (Violle *et al.* 2012).

Acknowledgements

We thank R. Allemand for *Leptopilina* populations and K. Varikou and V. Caleca for P. concolor samples (Crete and Sicily, respectively). We are also grateful to N. Ris for fruitful discussions, M. Thaon and N. Ris for rearing *Ceratitis capitata* and *Psyttalia* strains, and L. Kremmer for help with analysis of Leptopilina samples. This work was funded by the INRA Division 'Santé des Plantes et Environnement' (SPE), the French National Research Agency through CLIMEVOL ANR-08-BLAN-0231 and the 'Investments for the Future' LABEX SIGNALIFE: program reference ANR-11-LABX-0028, and the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 613678 (DROPSA). H. Mathé-Hubert Ph.D. was funded by the 'Provence Alpes Côte d'Azur' (PACA) region and the INRA SPE Division.

References

Aagaard JE, Yi X, MacCoss MJ, Swanson WJ (2006) Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. *Proceedings of the National Academy of Sciences, USA*, **103**, 17302–17307.

Alape-giron A, Sanz L, Madrigal M, Sasa M, Calvete JJ (2008) Snake venomics of the lancehead pitviper bothrops asper: geographic, individual, and ontogenetic variations. *Journal of Proteome Research*, **7**, 3556–3571.

Angulo Y, Escolano J, Lomonte B *et al.* (2007) Snake venomics of Central American pitvipers: clues for rationalizing the distinct envenomation profiles of *Atropoides nummifer* and *Atropoides picadoi*. *Journal of Proteome Research*, **7**, 708–719.

Armengaud J, Trapp J, Pible O *et al.* (2014) Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics*, **105**, 1–14.

Biostep (2008) Lane relationship studies. http://www.biostep.de/DocumentDownloadServlet?docid=48, 1–2.

Biron DG, Loxdale HD, Ponton F *et al.* (2006) Population proteomics: an emerging discipline to study metapopulation ecology. *Proteomics*, **6**, 1712–1715.

Blein-Nicolas M, Albertin W, Valot B *et al.* (2013) Yeast proteome variations reveal different adaptive responses to grape must fermentation. *Molecular Biology and Evolution*, **30**, 1368–1383.

Bobkov SV, Lazareva TN (2012) Band composition of electrophoretic spectra of storage proteins in interspecific pea hybrids. *Russian Journal of Genetics*, **48**, 56–61.

Bolnick DI, Amarasekare P, Araújo MS *et al.* (2011) Why intraspecific trait variation matters in community ecology. *Trends in Ecology & Evolution*, **26**, 183–192.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Burstin J, De Vienne D, Dubreuil P, Damerval C (1994) Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. *Theoretical and Applied Genetics*, **89**, 943–950.

Chevalier F, Martin O, Rofidal V *et al.* (2004) Proteomic investigation of natural variation between Arabidopsis ecotypes. *Proteomics*, **4**, 1372–1381.

Cheyppe-Buchmann S, Bon MC, Warot S *et al.* (2011) Molecular characterization of *Psyttalia lounsburyi*, a candidate biocontrol agent of the olive fruit fly, and its Wolbachia symbionts as a pre-requisite for future intraspecific hybridization. *BioControl*, **56**, 713–724.

Colinet D, Mathé-Hubert H, Allemand R, Gatti JL, Poirié M (2013) Variability of venom components in immune suppressive parasitoid wasps: from a phylogenetic to a population approach. *Journal of Insect Physiology*, **59**, 205–212.

Colinet D, Kremmer L, Lemauf S *et al.* (2014) Development of RNAi in a Drosophila endoparasitoid wasp and demonstration of its efficiency in impairing venom protein production. *Journal of Insect Physiology*, **63**, 56–61.

Crawford DL, Oleksiak MF (2007) The biological importance of measuring individual variation. *The Journal of Experimental Biology*, **210**, 1613–1621.

Dall SRX, Bell AM, Bolnick DI, Ratnieks FLW (2012) An evolutionary ecology of individual differences. *Ecology Letters*, **15**, 1189–1198.

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Davidson W (2012) Adaptation genomics: next generation sequencing reveals a shared haplotype for rapid early development in geographically and genetically distant populations of rainbow trout. *Molecular Ecology*, **21**, 219–222.

Diz AP, Skibinski DOF (2007) Evolution of 2-DE protein patterns in a mussel hybrid zone. *Proteomics*, **7**, 2111–2120.

Diz AP, Martínez-Fernández M, Rolán-Alvarez E (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology*, **21**, 1060–1080.

Domon B, Aebersold R (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology*, **28**, 710–721.

Dray S, Dufour AB (2007) The ade4 Package: implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22**, 1–18.

Dupas S, Poirié M, Frey F, Carton Y (2013) Is parasitoid virulence against multiple hosts adaptive or constrained by phylogeny? A study of Leptopilina spp./Drosophila interactions. *Annales de la Société Entomologique de France*, **49**, 222–231.

Fay JC, Wittkopp PJ (2008) Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, **100**, 191–199.

Feder ME, Walser J-C (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology*, **18**, 901–910.

Gonzalez EG, Krey G, Espiñeira M *et al.* (2010) Population proteomics of the European Hake (Merluccius merluccius). *Journal of Proteome Research*, **9**, 6392–6404.

Green-Church KB, Nichols KK, Kleinholz NM, Zhang L, Nichols JJ (2008) Investigation of the human tear film proteome using multiple proteomic approaches. *Molecular Vision*, **14**, 456–470.

Guimaraes JC, Rocha M, Arkin AP (2014) Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Research*, **42**, 4791–4799.

Hodgins-Davis A, Townsend JP (2009) Evolving gene expression: from G to E to GxE. *Trends in Ecology & Evolution*, **24**, 649–658.

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, **50**, 346–363.

Jacob AM, Turck CW (2008) Detection of post-translational modifications by fluorescent staining of two-dimensional gels. *Methods in Molecular Biology*, **446**, 21–32.

Jensen ON (2006) Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology*, **7**, 391–403.

Karam N, Guglielmino CR, Bertin S *et al.* (2008) RAPD analysis in the parasitoid wasp Psyttalia concolor reveals Mediterranean population structure and provides SCAR markers. *Biological Control*, **47**, 22–27.

Karl S, Toonen RJ, Grant WS, Bowen BW (2012) Common misconceptions in molecular ecology: echoes of the modern synthesis. *Molecular Ecology*, **21**, 4171–4189.

Karr TL (2008) Application of proteomics to ecology and population biology. *Heredity*, **100**, 200–206.

Karve TM, Cheema AK (2011) Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. *Journal of Amino Acids*, **2011**, 207691.

Kirk H, Freeland JR (2011) Applications and Implications of Neutral versus Non-neutral Markers in Molecular Ecology. *International Journal of Molecular Sciences*, **12**, 3966–3988.

Krishnan HB, Jang S, Baxter I, Wiebold WJ (2012) Growing location has a pronounced effect on the accumulation of cancer chemopreventive agent bowman-birk inhibitor in soybean seeds. *Crop Science*, **52**, 1786.

Laurent JM, Vogel C, Kwon T *et al.* (2010) Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics*, **10**, 4209–4212.

Li K-M, Ren L-Y, Zhang Y-J, Wu K-M, Guo Y-Y (2012) Knockdown of microplitis mediator odorant receptor involved in the sensitive detection of two chemicals. *Journal of Chemical Ecology*, **38**, 287–294.

López JL (2005) Role of proteomics in taxonomy: the Mytilus complex as a model of study. *Journal of Chromatography. B*, **815**, 261–274.

López J, Mosquera E, Fuentes J *et al.* (2001) Two-dimensional gel electrophoresis of *Mytilus galloprovincialis*: differences in protein expression between intertidal and cultured mussels. *Marine Ecology Progress Series*, **224**, 149–156.

Maier T, Güell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Letters*, **583**, 3966–3973.

Mathé-Hubert H, Gatti J-L, Poirié M, Malausa T (2013) A PCR-based method for estimating parasitism rates in the olive fly parasitoids *Psyttalia concolor* and *P. lounsburyi* (Hymenoptera: Braconidae). *Biological Control*, **67**, 44–50.

Melzer D, Perry JRB, Hernandez D *et al.* (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*, **4**, e1000072.

Morrissey JH (1981) Silver stain for proteins in polyacrylamide gels: a modified procedure with enhanced uniform sensitivity. *Analytical Biochemistry*, **117**, 307–310.

Mosquera E, López JL, Alvarez G (2003) Genetic variability of the marine mussel *Mytilus galloprovincialis* assessed using two-dimensional electrophoresis. *Heredity*, **90**, 432–442.

Navas A, Albar JP (2004) Application of proteomics in phylogenetic and evolutionary studies. *Proteomics*, **4**, 299–302.

Núñez V, Cid P, Sanz L *et al.* (2009) Snake venomics and antivenomics of *Bothrops atrox* venoms from Colombia and the Amazon regions of Brazil, Perú and Ecuador suggest the occurrence of geographic variation of venom phenotype by a trend towards paedomorphism. *Journal of Proteomics*, **73**, 57–78.

Papakostas S, Vasemägi A, Vähä J-P *et al.* (2012) A proteomics approach reveals divergent molecular responses to salinity in populations of European whitefish (Coregonus lavaretus). *Molecular Ecology*, **21**, 3516–3530.

Poirié M, Carton Y, Dubuffet A (2009) Virulence strategies in parasitoid Hymenoptera as an example of adaptive diversity. *Comptes Rendus Biologies*, **332**, 311–320.

Ramm S, McDonald L, Hurst JL, Beynon RJ, Stockley P (2009) Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Molecular Biology and Evolution*, **26**, 189–198.

Rees BB, Andacht T, Skripnikova E, Crawford DL (2011) Population proteomics: quantitative variation within and among populations in cardiac protein expression. *Molecular Biology and Evolution*, **28**, 1271–1279.

Reusch TBH, Wood TE (2007) Molecular ecology of global change. *Molecular Ecology*, **16**, 3973–3992.

Sanz L, Escolano J, Ferretti M *et al.* (2008) The South and Central American Bushmasters. Comparison of the toxin composition of *Lachesis muta* gathered from proteomic versus transcriptomic analysis. *Journal of Proteomics*, **71**, 46–60.

Schrimpf SP, Weiss M, Reiter L *et al.* (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biology*, **7**, e48.

Sickmann A, Marcus K, Schäfer H *et al.* (2001) Identification of post-translationally modified proteins in proteome studies. *Electrophoresis*, **22**, 1669–1676.

Slattery M, Ankisetty S, Corrales J *et al.* (2012) Marine proteomics: a critical assessment of an emerging technology. *Journal of Natural Products*, **75**, 1833–1877.

Smyth GK (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S), pp. 397–420. Springer, New York, New York.

Smyth GK, Yang YH, Speed T (2003) Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology*, **224**, 111–136.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Valcu C-M, Kempenaers B (2014) Proteomics in behavioral ecology. *Behavioral Ecology*, **26**, 1–15.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, 4th edn. Springer Verlag, New York, New York.

Violle C, Enquist BJ, McGill BJ *et al.* (2012) The return of the variance: intraspecific variability in community ecology. *Trends in Ecology & Evolution*, **27**, 244–252.

Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M (2011) Regulatory variation within and between species. *Annual Review of Genomics and Human Genetics*, **12**, 327–346.

T.M., M.P., J.-L.G., H.M.-H. and D.C. conceived and designed the method. H.M.-H. wrote R codes and performed data analysis. H.M.-H. performed the experiments. J.-L.G., T.M., M.P. and H.M.-H. wrote the manuscript.

Data Accessibility

R scripts and functions and help files: see online supporting information (Appendix S1) or the zip file attached to this pdf file. Data sets supporting the results of this article are available in the Dryad repository http://datadryad.org/review?doi=doi:10.5061/dryad.d1d1m.

Supporting Information

Additional Supporting Information may be found in the following pages:

Appendix S1 R scripts and functions, with help files.

Appendix S2 General statistical and practical advices and possible improvements of the method.

Fig. S1 Schematic diagram of the main analysis steps.

Fig. S2 Example of graph of a median profile and reference bands produced by the (*plot.profile*) function.

Fig. S3 Sensitivity experiment results: saturation range and gels used to test the linearity between the 'sensitivity to staining duration' and the 'sensitivity to protein quantity'.

Fig. S4 Sensitivity experiment results: Assessment of the power of the method to describe variation in band intensity.

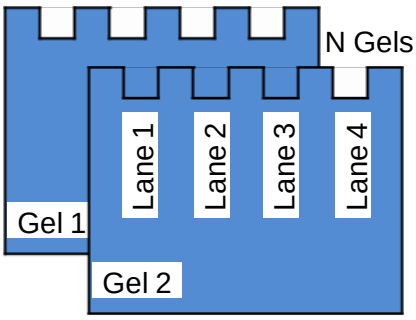
Fig. S5 Examples of the gels used in case studies and identification of band #6.

Table S1. Summary of the four gamma GLM fitted to the ratio of inter group/ intra group variability.

Figure S1.
Details of the main analysis steps

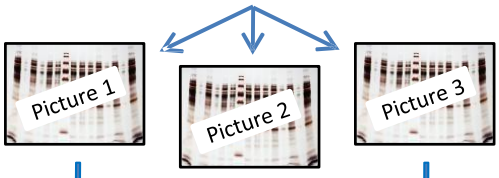
The two first steps summarize the experimental design, the third step correspond to the pre-processing of gel pictures with the Phoretix software. The next steps were performed using the developed R functions.

Several gels



Several pictures for each gel

Several pictures
by gel



Picture analysis

For each picture: lane alignment
within and between gels with Rf lines

For each picture: Lane profile exportation
with AND without the background

Phoretix

Profile analysis

without background
with background

For each profile (each lane and picture):
Band detection and quantification (with height and volume)

For each lane:
Picture selection

Choice of dark pictures for lanes with low amounts of material
Choice of light pictures for lanes with high amounts of material

Data analysis:
Normalization

Statistical normalization for the remaining
heterogeneity in material amount between lanes
(three normalization methods are tested)

Comparison of
methods

Comparison of results :
With or without the background
With the height or the volume
With each of the three normalization methods

Choice of one combination of
methods

R

Data set

	Gel 1 Lane 1	Gel 1 Lane 2	. . .	Gel 2 Lane 1	Gel 2 Lane 2	. . .
Band1						
Band2						
. . .						
Band ...						

Normalised intensities

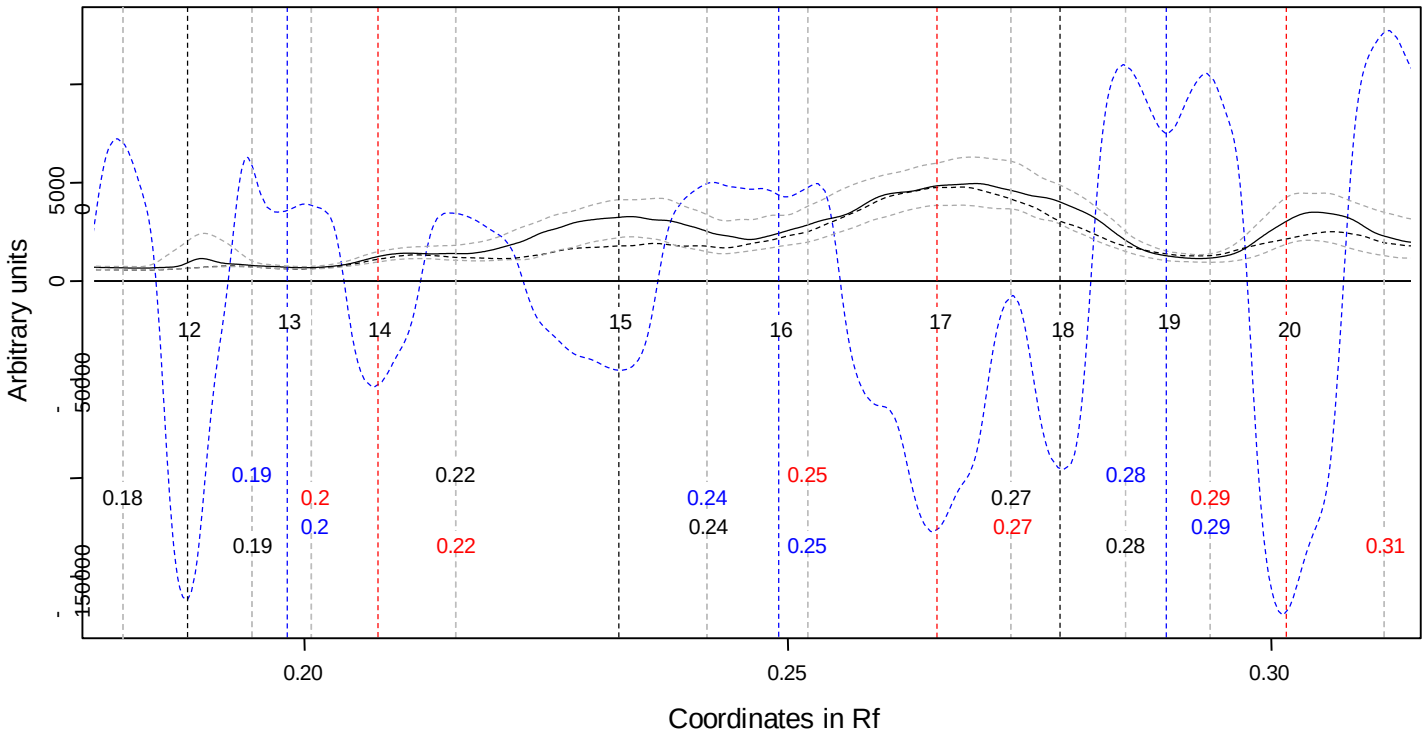


Figure S2. Example of a graph produced by the *plot.profile* function

The graph shows (i) the median profile and part of its variability (ii) the second derivative that is used to detect the peaks and (iii) the automatically detected peaks. These peaks have to be edited manually, this task being performed using both the gel pictures and the graphs. Due to space constraints, only part of the whole Rf range (0 to 1) is shown. The solid and dotted black lines represent the weighted and unweighted median of intensities, respectively. Dotted grey lines are the weighted quartiles of median. The dotted blue line, the most informative, corresponds to the weighted median of the second derivatives used for the automatic detection of reference peaks. Colored vertical lines (by default, blue, red and black) are the peak positions (local minima of the weighted median of second derivatives), and grey vertical lines are the borders of the peaks (local maxima of the weighted median of second derivatives). Black numbers below the horizontal 0 line are the ID (reference number) of peaks, which allow the handling of peaks in the function *modify.peaks.manually*. Colored numbers (on grey dotted lines) indicate the Rf coordinates of the borders of peaks. The color of these numbers correspond to the color of the vertical line that indicates the position of the center of the peak to which borders coordinates refer. Colored numbers are positioned on four lines on the y axes, the two first ones corresponding to left borders, the others to right borders.

Peaks shown on the figure have been automatically detected and manual corrections have to be performed using the function *modify.peaks.manually*. For example, peak 19 may not exist at all, while supplementary peaks are likely present between peaks 14 and 15 and peaks 16 and 17 (to be confirmed on gels pictures). Band 12 is typically a rare but intense band as the weighted 0.75 quartiles of median (grey dotted line) is mutch higher than the weighted median (solid black line).

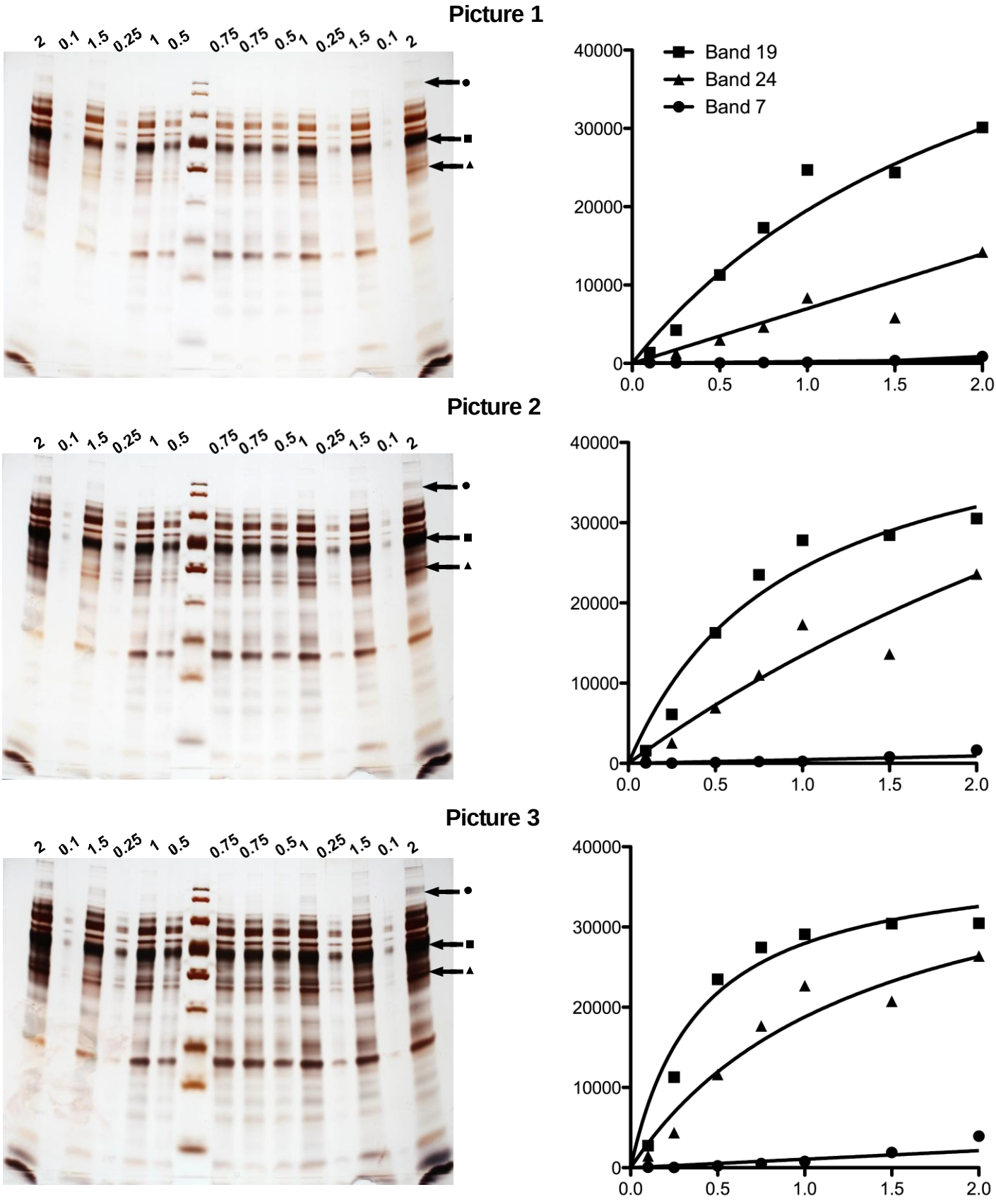


Figure S3. Sensitivity experiment results: Saturation range

Pictures 1, 2 and 3 correspond to the three pictures taken at different times during the gel staining procedure. Values above the pictures indicate the amount of loaded sample as individual venom gland fractions. Plots illustrate the relationship between the amount of loaded sample and the intensity of three bands (7, weak intensity - 19, high intensity - 24, medium intensity), indicated on gels by arrows.

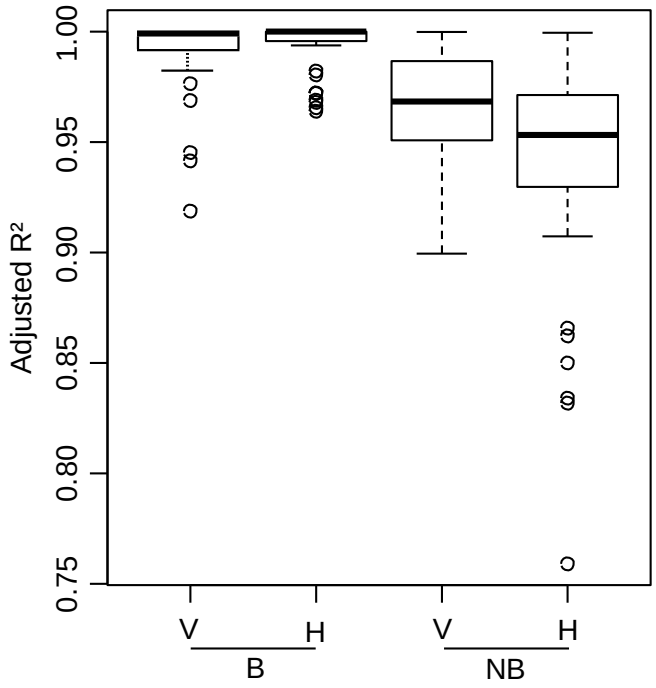


Figure S4. Sensitivity experiment results: Assessment of the power of the method in describing variation in band intensity.

The boxplots represent for each reference band and each of the four parameter combinations, the ajusted-R^2 of models explaining the band intensity by the protein quantity, the staining duration and their interaction. “V” and “H” letters indicate respectively if the peak volume or the peak height was recorded. “B” and “NB” letters correspond to the analysis of profiles with or without the background, respectively.

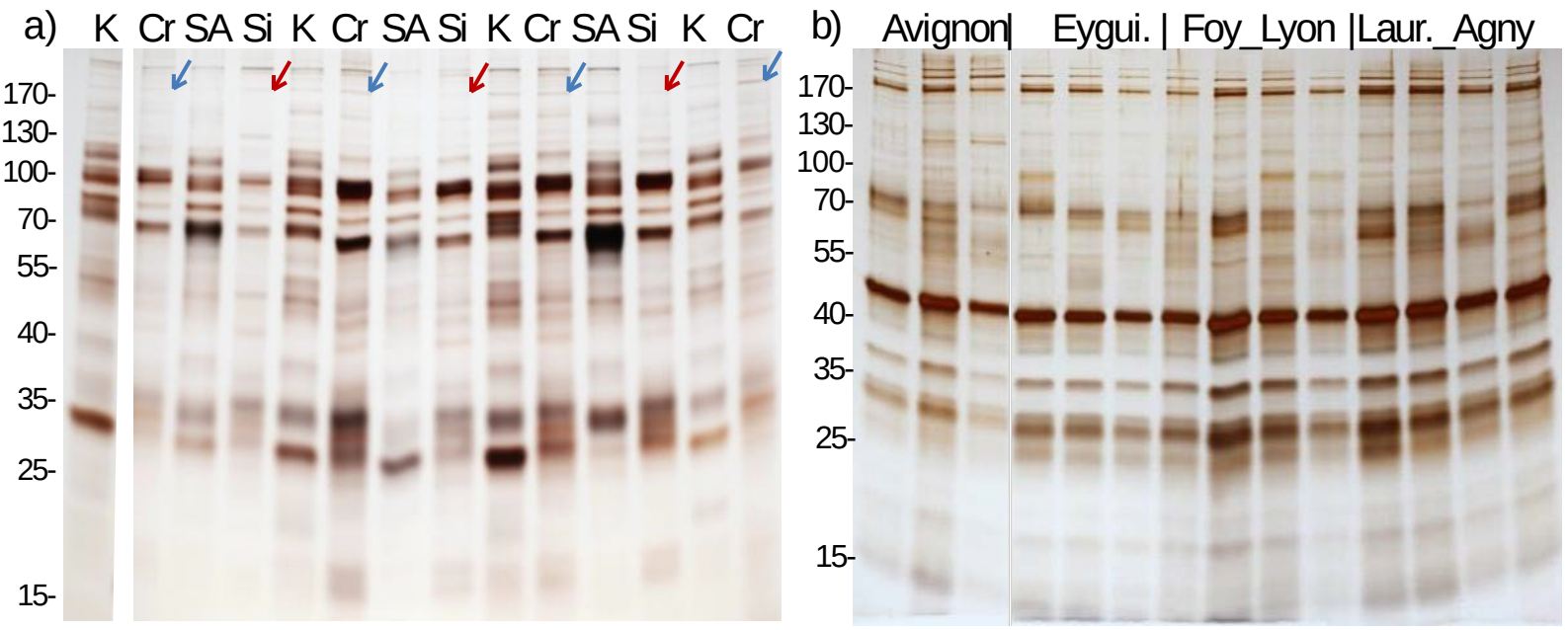


Figure S5. Examples of the gels used to test the analysis method

a) Picture of a gel used to analyze *Psyttalia* spp. venom. Letters indicate the species and population for each individual. “K”, Kenyan *P. lounsburyi*; “SA”, South African *P. lounsburyi*; “Cr”, Cretan *P. concolor*; “Si”, Sicilian *P. concolor*. Arrows points to the *P. concolor* band #6 which variability in intensity is described Figure 4c. (blue and red arrows for Cretan and Sicilian individuals, respectively). b) Picture of a gel used to analyze *L. boulardi* venom. Molecular weight standards are in kDa.

Table S1. Summary of the four gamma GLM fitted to the ratio of inter group / intra group variability

Groups	Explanatory variables	Likelihood-ratio test		
		χ^2	d.f.	*P value*
Psyttalia sp (*P. lounsburyi*, *P. concolor*)	Parameters	38.18	11	$< 1.10^{-3}$
	Ref Bands	571.67	37	$< 1.10^{-3}$
P. concolor (Crete, Sicily)	Parameters	39.85	11	$< 1.10^{-3}$
	Ref Bands	388.31	37	$< 1.10^{-3}$
P. lounsburyi (South Africa, Kenya)	Parameters	34.33	11	$< 1.10^{-3}$
	Ref Bands	713.80	37	$< 1.10^{-3}$
L. boulardi (Avignon, Eyguières, Ste Foy Lès Lyon, St Laurent d'Agny)	Parameters	34.26	11	$< 1.10^{-3}$
	Ref Bands	380.62	31	$< 1.10^{-3}$

Gamma GLM models are used to compare, for each case study ("Groups"), the power of the parameters combination in detecting the inter-group structure. For each model, the explained variable is the ratio between the inter group and the intra group variability.

Appendix S2: General statistical and practical advices

***In silico* sample quantities homogenization by the use of several pictures (*select.photo*)**

Sample quantities can be homogenized *in silico* using pictures of different staining levels.

To this end, the function (*select.photo*) selects, for a given lane, the picture ensuring the best match between the intensity of the lane and the median[1] intensity of all lanes (of all gels).

It is recommended that each lane has one picture with the desired coloration level. Thus the number of pictures to analyze for a given gel will depend on the level of variability in protein quantities on the gel. It is advised to take many pictures during coloration (e.g. 10 or even more pictures), and then to choose a subset of pictures to analyze.

What before and after this method?

In terms of data analysis, the method developed is positioned between two major steps of the 1D data analysis. Upstream, it requires the preliminary conversion of gel pictures into a set of aligned intensity profiles (second step of the Fig. S1).

For some kind of profiles (e.g. gels of low complexity or variability), future improvement may come from the automation of the alignment between lanes, instead of manual setting of Rf lines (as we did with Phoretix). This might be done using the R package PTW (Parametric Time Warping; Eilers 2004) that uses aligned profiles using warping for optimizing cross-correlation between profiles. But this seems to be only efficient if the cross-correlation between profiles (once aligned) is high, or if peaks are well delimited.

Downstream, statistical analyses can be performed on the output of our procedure, which is composed of a great number of continuous variables, each corresponding to a reference band.

---

[1]: or other manually set value

Below, we present a list (far from exhaustive) of the statistical tools that could be useful to analyze the datasets provided by the method. These datasets are characterized by the presence of many continuous variables, and multivariate statistics are thus the most suitable for their analysis.

Generally, the first step will be to describe the main features of the dataset. To this end, methods such as PCA that allows reducing the dimensionality of datasets, or heatmaps that allow the sorting of individuals into clusters and the characterization of these clusters by clusters of bands may be useful. R packages dedicated to the analysis of ecological data are specifically suitable to this aims. The ADE4 package implements many useful multivariate tools (such as within and between PCA, instrumental PCA, and others K-tables methods), a large documentation is available on the website (Dray & Dufour 2007; Dray *et al.* 2007), and this package contains a Graphical User Interface. However, many other packages can also be successfully used (e.g.: "labdsv", "cluster", "pvclust", and "vegan"; Dixon 2003; Suzuki & Shimodaira 2006; Roberts 2013; Maechler *et al.* 2014). For example, the "pvclust" package allows testing the significance of a given cluster viewed in a heatmap.

An objective of the analysis performed with the method might be to test the effect of the variation of protein composition on a particular phenotype. This approach requires first to test if the overall protein variability has an effect on the phenotype of interest. This can be achieved using multivariate regression, possibly with a previous reduction of dimensionality through PCAs methods. If the overall protein variability has a phenotypic effect, a next step would be to identify bands that contain proteins involved. This will be particularly difficult if bands of interest are prone to multcollinearity, a recurrent problem in studies aiming to identify such candidates. Indeed, accurate identification of the respective roles of highly correlated factors requires large sample sizes. Various statistical

methods allowing to handle multcollinearity have been compared and discussed by Dormann *et al.* (2013) and El-Dereny & Rashwan (2011).

But the level of multicollinearity can also be modulated by the choice of the combination of parameters. For instance, removal of the background (mainly linked to the problem of adjacent bands), will reduce multicollinearity. Similarly, although the volume may be more accurate than the height to quantify band intensity, it measures intensity at the borders of the band and thus is more likely to co-vary with adjacent bands. **This should be taken into account when choosing a combination of parameters.**

Another approach might be to virtually merge too highly correlated bands in the analysis and accept the uncertainty about the respective role of the proteins they contain. This approach may be particularly valuable if few bands are highly correlated since this will only increase the number of candidate proteins to investigate. Importantly, it is also expected that some bands are highly correlated for biological reasons. For example, bands containing subunits of a same protein are expected to be strongly positively correlated. An interesting and friendly discussion about multicollinearity can be found here: http://psychologicalstatistics.blogspot.fr/2013/11/multicollinearity-and-collinearity-in.html.

Another aim of such analyses might be to test if geographical or environmental factors may explain multivariate differences in the protein expression MANOVA would be particularly suitable for such analyses but it requires multivariate normality that is sometimes impossible to obtain. The non parametric MANOVA (Anderson 2001) that is implemented by the "adonis" function of the R package "vegan" could be a good alternative. Interestingly, this approach has already been used for proteomic analysis (Zerzucha *et al.* 2012). However, caution must be taken when interpreting the results of this function (Roff *et al.* 2012b). Functions developed by Roff *et al.* (2012a) and Aguirre *et al.* (2014) might be more appropriate.

Many of these tools are available in other statistical software.

Applicability of the method for the analysis of bands on Western-blots

Although the method can easily be used for Western-blot analysis, the normalization step must be avoided because of the very low number of bands usually analyzed. In addition, the interest of the method is that it allows a very rapid analysis of a large number of bands. It may be less useful when a low number of bands are to be analysed.

References

Aguirre JD, Hine E, McGuigan K, Blows MW (2014) Comparing G: multivariate analysis of genetic variation in multiple populations. *Heredity*, **112**, 21–9.

Anderson M (2001) A new method for non- parametric multivariate analysis of variance. *Austral Ecology*, 32–46.

Dixon P (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, **14**, 927–930.

Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dray S, Dufour AB (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal Of Statistical Software*, **22**, 1–18.

Dray S, Dufour AB, Chessel D (2007) The ade4 Package — II: Two-table and K -table Methods. *R News*, **7**, 47–52.

Eilers PHC (2004) Parametric time warping. *Analytical Chemistry*, **76**, 404–11.

El-Dereny M, Rashwan N (2011) Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Science*, **6**, 585–600.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2014) cluster: Cluster Analysis Basics and Extensions.

Roberts DW (2013) labdsv: Ordination and Multivariate Analysis for Ecology.

Roff D a, Prokkola JM, Krams I, Rantala MJ (2012a) There is more than one way to skin a G matrix. *Journal of Evolutionary Biology*, **25**, 1113–1126.

Roff DI, Wright ST, Wang Y (2012b) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)*, **22**, 1540–2.

Zerzucha P, Boguszewska D, Zagdańska B, Walczak B (2012) Non-parametric multivariate analysis of variance in the proteomic response of potato to drought stress. *Analytica Chemica Acta*, **719**, 1–7.