MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 3223–3231

NEWS AND VIEWS

OPINION

Seven common mistakes in population genetics and how to avoid them

PATRICK G. MEIRMANS

Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, P.O. Box 94248, 1090GE Amsterdam, Netherlands

As the data resulting from modern genotyping tools are astoundingly complex, genotyping studies require great care in the sampling design, genotyping, data analysis and interpretation. Such care is necessary because, with data sets containing thousands of loci, small biases can easily become strongly significant patterns. Such biases may already be present in routine tasks that are present in almost every genotyping study. Here, I discuss seven common mistakes that can be frequently encountered in the genotyping literature: (i) giving more attention to genotyping than to sampling, (ii) failing to perform or report experimental randomization in the laboratory, (iii) equating geopolitical borders with biological borders, (iv) testing significance of clustering output, (v) misinterpreting Mantel's *r* statistic, (vi) only interpreting a single value of *k* and (vii) forgetting that only a small portion of the genome will be associated with climate. For every of those issues, I give some suggestions how to avoid the mistake. Overall, I argue that genotyping studies would benefit from establishing a more rigorous experimental design, involving proper sampling design, randomization and better distinction of a priori hypotheses and exploratory analyses.

Keywords: AMOVA, clustering, genome scan, Mantel test, population structure, unicorns

Received 23 February 2015; revision received 1 May 2015; accepted 11 May 2015

Introduction

Genetic data sets are rapidly getting bigger. Along with this increase in size, there has also been a broadening of the scope of genotyping studies. They are no longer only used to answer purely genetic questions, but increasingly also to answer ecological questions (as exemplified by the name of this journal). This increase in scope is exciting as, for the first time, we are starting to get the whole picture of how selection works: from the genetic variation up to the individual phenotype, the population and the environment. On the other hand, the ease with which genetic data can be obtained attracts many researchers from other fields, who lack a formal training in population genetics (Karl *et al.* 2012).

Large data sets present many challenges (Karl *et al.* 2012; Vrijenhoek & Waples 2012), among which is that small mistakes in experimental design or data analysis may lead to false inferences. A very small bias may have a negligible effect in a microsatellite data set, but become a strongly significant pattern in a next-generation sequencing project. Sometimes, such biases are recognized and controlled for, but frequently they lead to false positives and incorrect inferences (Lotterhos & Whitlock 2014). Such problems already appear in very simple types of analyses: recently, Waples (2014) pointed out some common mistakes made in testing and interpreting Hardy–Weinberg equilibrium, a concept that is more than a century old. However, there are more issues where we, as the scientific community, have gradually slipped into the habit of doing relatively simple things incorrectly.

In this study, I discuss seven mistakes that are frequently made in genotyping studies and that greatly hinder our ability to assess which processes determine the distribution of genetic variation. These issues are not necessarily all limited to modern genomic data sets, but may also occur with smaller data sets such as microsatellite studies. However, the size of modern data sets makes the issues more pertinent, as their sheer power will lead to very small *P*-values, making researchers overly confident that the patterns they are seeing are real. For every issue, I provide some simple suggestions on how these types of problems can be avoided. I present these issues in the order in which a typical genotyping study is performed: sampling, genotyping, analysis and finally the interpretation of the results.

Common mistakes

Giving more attention to genotyping than to sampling

In a recent review on how we can use genetics to find local adaptation, Tiffin & Ross-Ibarra (2014) included a large section called 'challenge of obtaining high-quality data'. This section discussed the difficulties and pitfalls in obtaining high-throughput sequencing data, but they did not say anything about the sampling strategy. This is a major omission as high-quality data start with high-quality sampling. The choice of sampling strategy may even have a stronger impact on the outcome of a genotyping study than the peculiarities of the marker system. High-throughput sequencing projects have—for the time being—generally

Correspondence: Patrick G. Meirmans, Fax: +31 20 5257832; E-mail: p.g.meirmans@uva.nl

lower sample sizes than classical genotyping studies, making it more important where exactly the included individuals come from. A good sampling strategy can be expensive, especially for widespread and/or elusive species. Therefore, faced with limited financial resources, researchers often prefer to spend their money on additional genotyping than on sampling. This is unfortunate as a failure to invest in robust sampling may completely waste the investment in genotyping.

Sampling design can have large effects on the outcome of commonly used population genetic analyses. A simple and well-known example is provided by gaps in the sampling that can lead to a false detection of population clustering when there is in fact a simple pattern of isolation by distance (IBD) (Audzijonyte & Vrijenhoek 2010). In tests of correlations between genetic variation and environmental variables (genetic–environment associations, GEA; Hedrick *et al.* 1976), linear transects across a strong environmental gradient are wholly uninformative as they only present a single axis of variation that will overlap with neutral processes (Meirmans 2012a). In designing a sampling strategy, there is often a trade-off between sampling few populations with many individuals and sampling many populations with few individuals. Different statistical methods differ in how their power is affected by this trade-off. For example, assignment tests depend on accurate estimates of population allele frequencies and work best with few population samples of many individuals (Meirmans 2012a). On the other hand, for GEA tests, many populations with few individuals are often preferable, because that will maximize the environmental variance (Lotterhos & Whitlock 2015). Because of these trade-offs, it is unwise to perform both types of tests in a single study as at least one of them will have insufficient power. In a comparable way, different analyses may require different spatial scales of sampling.

The particularities of a species' distribution may also limit the possible sampling designs. This can then lead to pseudoreplication in the sampling that is difficult to account for. In many cases, especially when rare species are concerned, it may therefore actually not be possible to have a meaningful sampling at all when one wants to do a certain analysis. Imagine fictitious montane unicorns (*Monoceros montanus*; Hurlbert 1990) that have a distribution consisting of two geographically and genetically well-separated areas (North and South), with only a small number of demes in each. In such a case, both the main environmental variance and the main genetic variance lie between the two areas, and their covariance makes analyses of GEA nearly impossible.

What to do. Design your sampling based on simulations specific to your study species. Many excellent tools are available for doing population genetic simulations (Hoban *et al.* 2012), some of which can simulate realistic geographical scenarios (Meirmans 2011). For many species, distribution data are available in public databases (e.g. gbif.org) that can be used to fit the simulated metapopulation to the study species. The simulated data can then be used (possibly together with available environmental data) to assess the performance of different types of analyses under different sampling strategies. For GEA tests, paired sampling of nearby patches with opposing environments seems to provide the highest power (Lotterhos & Whitlock 2015), but this may depend on the spatial distribution of the species and on the environmental variables. Finally, be willing to acknowledge defeat: not all types of analyses may be meaningful for your target species. In those cases, be prepared to skip the analyses all together.

Failing to perform or report experimental randomization in the laboratory

Let us assume that we want to test whether male and female montane unicorns differ in the effect of diet on the length of their horns. In such an experiment, we should not place all male foals in one stable and all females in another. Instead, the two sexes should be randomized over stables in order to prevent bias due to possible effects of stable on horn length. Randomizations are so essential to experimental studies that most journals refuse to publish studies that lack a proper description of how exactly subjects were randomized over treatments. Strangely, however, genotyping studies seem to be exempted from this rule. Few genotyping papers describe in full detail whether and how individuals and populations were randomized during the genotyping analysis in the laboratory.

Genotyping studies are in fact experiments. Therefore, they should be executed in a fashion that prevents bias stemming from the genotyping process: for example, if we want to test for genetic differences among populations, we should employ proper randomization of individuals over populations to prevent bias. There is plenty of room for bias. Even though we like to think of genetic data as very solid, anybody who has spent some time in the laboratory knows that this is not the case. With AFLPs, differences in the performance of the PCR may cause differences in band intensity and therefore lead to differences in genotypes when gels are scored. With microsatellites, most types of errors may be gel or plate specific, including stutter bands, ghost bands, mistakes in the estimation of band length and other artefacts. With next-generation sequencing studies, the coverage, and hence the degree of missing data, may vary strongly among lanes and there may be highly replicable errors in base-calling. When all individuals from one population are placed on one plate, gel or lane, and the individuals from another population on another, errors that are specific to a plate, gel or lane will lead to an overestimation of the among-population differentiation, the presence of artificial clusters and an underestimation of the migration rate.

What to do. As in every experiment, proper randomization should be employed during the genotyping in the laboratory. At the most basic level, individuals from different populations should be randomized over gels. In studies of

spatial autocorrelation, you should also prevent that individuals or populations that are geographically close in nature are also close to each other on the genotyping plate, gel or lane; the randomization of such data should therefore take the geographical coordinates into account. It is perfectly possible that such randomization is already practised in genotyping laboratories everywhere and I am simply unaware of it. However, if this is the case, this is nowhere evident in the literature. Therefore, editors and reviewers should make sure that studies clearly report what measures have been taken to prevent such biases.

Equating geopolitical borders with biological borders

Detecting transitions in the distribution of genetic variation is an important aspect of many genotyping studies, be it to make inferences about past historical events (Taberlet *et al.* 1998) or to correct for biases it may induce (Excoffier *et al.* 2009; Meirmans 2012a). Often, this is performed by hypothesizing the location of the border a priori and then testing its significance (e.g. in an AMOVA). This approach gives a good inference power as it is based on the direct testing of an hypothesis, in contrast to more exploratory types of analyses such as STRUCTURE (Pritchard *et al.* 2000). To obtain such an inference power, there must therefore be a strong, biologically valid, a priori, hypothesis. However, what we often see is that a priori groups are specified that lack a solid biological reasoning.

In practice, the groups are often made by simply dividing the sampled area into, for example, an 'eastern' and a 'western' cluster. Such groupings can even be based on purely anthropogenic factors, such as geopolitical borders. It is obvious that there is very little insight to be gained from such an analysis as wild species are hardly ever concerned with geopolitical borders. Any effects of geopolitical borders on dispersal are only expected in cases where the border coincides with a major geographical border, such as a sea, mountain range or river. In rare cases, and mostly for larger animals, there may also be a barrier when the border is closed to such a degree to become a major impediment to dispersal (e.g. the former iron curtain, the Korean demilitarized zone and the US–Mexico barrier). Even so, in the latter cases, the border may be too recent to have left any biologically relevant effects on the genetic population structure.

What to do. Rather than concentrating on ad hoc or geopolitical borders, only use tests of a priori specified structure when you have a sound biological hypothesis about the location of the border. If you have a general idea, but are unsure about the exact location of a border, use a (spatial) clustering algorithm instead (Pritchard *et al.* 2000; Jombart *et al.* 2010; Meirmans 2012b). Note that, especially with large genomic data sets, AMOVAs are powerful enough to also give strongly significant results when the defined structure is only partly correct, which may give you false confidence. For example, it is well described that many European species show an East–West clustering, resulting from postglacial recolonization (Taberlet *et al.* 1998). However, the locations of the genetic transition zones vary widely among species. Therefore, testing a generic 'East–West' distinction for a particular species will not offer any additional insights.

Testing significance of clustering output

Let us assume we have a sample of montane unicorns and have measured for each individual the length of its horn and then separated the individuals into two groups based on their horn length. Anyone with a passing knowledge of statistics knows that we are then not allowed to test whether the long-horned and short-horned unicorns have different horn lengths. This is because there is an obvious case of nonindependence as the grouping was based on horn length to begin with. The same argument holds for multivariate data: if clusters have been assigned—either manually or algorithmically—based on a particular data set, we are not allowed to test whether the clusters are significantly different for this same data set.

The same rule should be applied to genotyping data, because that is essentially multivariate. Nevertheless, this practice is surprisingly common in population genetic studies: first, a clustering algorithm, such as STRUCTURE (Pritchard *et al.* 2000), DAPC (Jombart *et al.* 2010) or K-means (Meirmans 2012b), is used, and the optimal number of groups is established. Subsequently, an AMOVA (Excoffier *et al.* 1992) is performed based on these groups, including—and wrongly so—a permutation test to assess the significance of the differentiation. The *P*-value resulting from such a permutation test is completely meaningless as the test suffers from exactly the same circularity as the unicorn example above. When the clustering is carried out using STRUCTURE, the problem is especially egregious because that represents a mixture of statistical frameworks: the Bayesian STRUCTURE results are being retrofitted into the probabilistic permutation test.

What to do. You can easily avoid this problem by not reporting the *P*-values of an AMOVA that is based on clustering of the same data. Note that it is perfectly fine to report the *F*-statistics from such an AMOVA; these give an indication of the strength of the separation between the clusters, which may be valuable information for the interpretation of the results. *P*-values, on the other hand, do not measure the strength of differentiation, but only indicate the level of statistical support. In such a case, the *P*-values—which may be strongly significant for a NGS data set—suffer from circular reasoning and are therefore meaningless.

Only interpreting a single value of k

Clustering data is difficult. Therefore, many different algorithms have been developed, each with a different approach and different assumptions (e.g. Pritchard *et al.* 2000; Francois *et al.* 2006; Jombart *et al.* 2010; Meirmans 2012b). Determining the number of clusters (k) is even

harder. This is partly because increasing the number of clusters will generally improve the fit of the model on the data. A proper method therefore includes a correction for this, although different data sets may require a different correction. Back in 1985, Milligan & Cooper found no less than 30 procedures to estimate the number of clusters, and many more procedures have been added since. Nowadays, the most popular clustering method for genetic data is STRUCTURE (Pritchard *et al.* 2000), where two summary statistics are used for inferring *k*: Ln P(D) (Pritchard *et al.* 2000) and ΔK (Evanno *et al.* 2005). Interestingly, these statistics have been described by their respective authors as 'dubious at best' and 'ad hoc'. This indicates that generally there is a large degree of uncertainty in estimates of *k* and that these should be taken with a generous helping of salt.

The demographic, environmental and historical processes that have led to the current genetic structure of species are multifaceted and complex. As a result, there may be different levels of organization present in the genetic structure. In many studies, a simple hierarchical structure is assumed where individuals are clustered within populations and populations are clustered within geographical regions. In such a simple structure, there are already two valid values of *k*: the number of populations and the number of regions. However, the populations within regions may be variable in their rates of gene flow and their shared ancestry, such that some populations may stand out more than others, leading to more complex patterns of population structure. Figure 1 shows an example using simulated data of a system with a tree-like structure (see Betto-Colliard *et al.* 2015 for an example from real data). Although ΔK gives an optimum of two clusters for this data set, we see that values of *k* from 2 to 5 are all informative about the actual population structure. In nature, the different levels also do not necessarily have to be hierarchically structured: a species may be structured into geographical regions, but across regions also into ecotypes or host races. Therefore, we see that, besides the large uncertainty in the estimation of *k*, there often are few biological reasons to assume only a single value of *k*. Different values may simply reflect different demographic processes and therefore warrant interpretation.

What to do. Discuss all clustering results that warrant a biological interpretation. A clustering analysis is by nature a type of exploratory analysis, which makes it more open to interpretation at multiple levels; STRUCTURE is actually used in such a way in the original paper describing the method (Pritchard *et al.* 2000). Of course you should also observe some prudence to prevent overinterpretation: only patterns with a clear biological explanation should be considered. This also applies to your interpretation of the *k*-value that is deemed optimal according to the summary statistic of choice: it is better to have a biologically interpretable pattern from a 'suboptimal' *k* than a completely

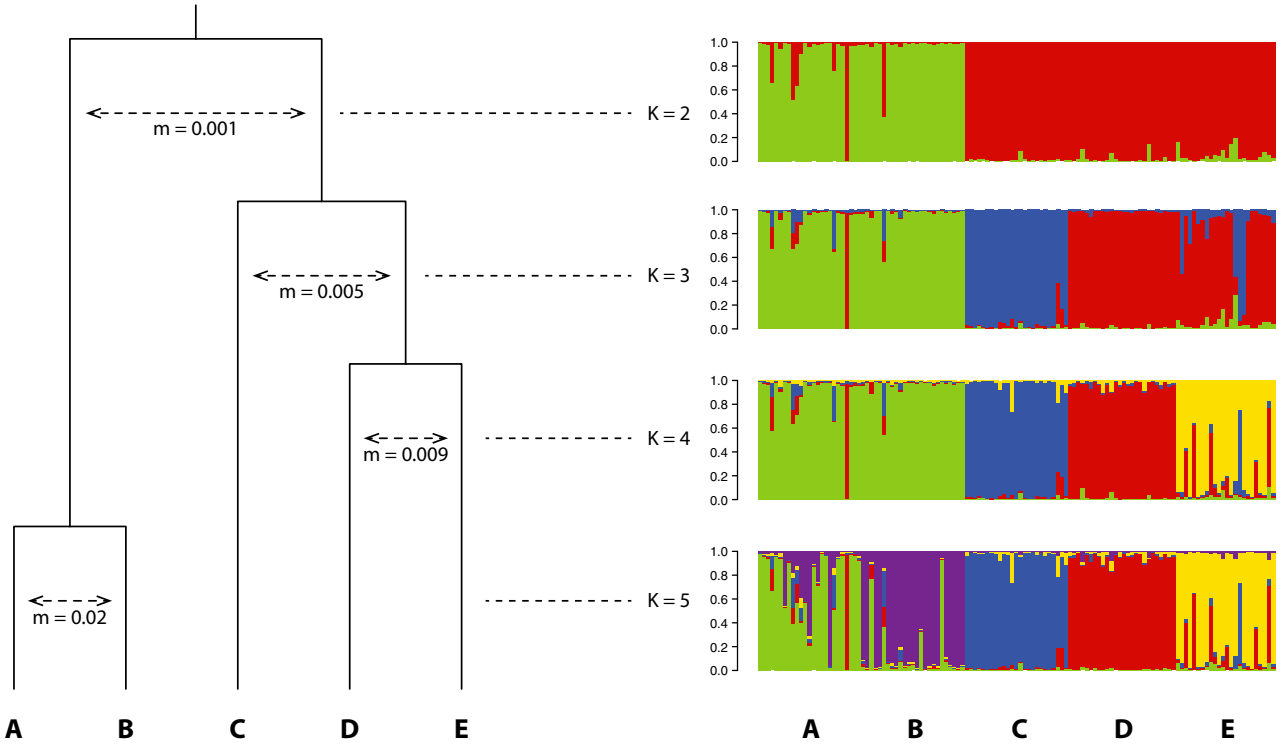

Fig. 1 STRUCTURE results (right) for simulated data under a tree-like population structure (left). Although the optimum number of clusters returned by ΔK is *k* = 2, all levels of *k* presented here are informative about the actual population structure. The migration rates among the clusters are indicated by the arrows in the tree. This figure was inspired by Fig. 2 in Betto-Colliard *et al.* (2015). See the Supporting information for more information regarding the simulation and the analysis of the simulated data.

unrealistic pattern from the 'optimal' k. In contrast, there are also cases where there is a clear a priori expectation about k; in such cases, it may be unnecessary to interpret multiple values of k. This is, for example, the case when explicitly analysing the hybridization and admixture between a number of well-established species (e.g. Van Hengstum *et al.* 2012).

Misinterpreting Mantel's r statistic

In most species, if not all, dispersal is spatially limited and most offspring disperses only a relatively short distance from its parents. In such cases, we say that there is IBD, as demes that are close together tend to exchange more genes than demes that are far apart. Testing for IBD is therefore a standard analysis in population genetic studies; this is typically done by performing a Mantel test (Mantel 1967) on a matrix of pairwise genetic distances (e.g. linearized F_{ST}) and a matrix of pairwise geographical distances. The Mantel test is a very flexible tool, as many ecological hypotheses can be phrased as an association between pairwise matrices (Legendre & Legendre 1998). For example, it is also frequently used to test for isolation by adaptation (IBA), or to compare IBA and IBD (Shafer & Wolf 2013). Unfortunately, the interpretation of Mantel's r statistic is far from straightforward in these cases.

The main problem is that Mantel's r (or the r^2 from a linear regression on the matrices) does not provide an accurate decomposition of the genetic variation (Legendre & Fortin 2010). Instead, it gives a decomposition of the variation in genetic distances, although it is also not very good at that (Raufaste & Rousset 2001). As a result, Mantel's r does not say anything about what we are actually interested in when we use a Mantel test: how the spatial distribution of the genetic variation within a species is affected by dispersal limitation.

Let us first derive an expectation of how the degree of spatial autocorrelation in the total genetic variance is influenced by the migration rate. When the migration rate is very low, we obtain high values of F_{ST}, indicating that almost all genetic variation lies among populations. Because of the extremely limited dispersal, we expect that only a small proportion of the total variation is spatially autocorrelated (Lotterhos 2012; Puebla *et al.* 2012). On the other extreme, when migration is very high, F_{ST} is low and almost all variation lies within populations. In this case, there again is very little of the total genetic variation left to be spatially autocorrelated. So we expect that the fraction of the total genetic variation that is spatially autocorrelated reaches its maximum value at intermediate migration rates. To illustrate the limitations of Mantel's r in this respect, I performed some simulations of a simple linear metapopulation with stepping-stone migration, using the software MARLIN (Neuenschwander *et al.* 2008; Meirmans 2011; see Supporting information data for details). Figure 2 shows that Mantel's r does not display the expected pattern, but instead increases monotonically with increasing migration rate. Therefore, a high value of Mantel's r does not indicate limited dispersal, but rather the opposite.

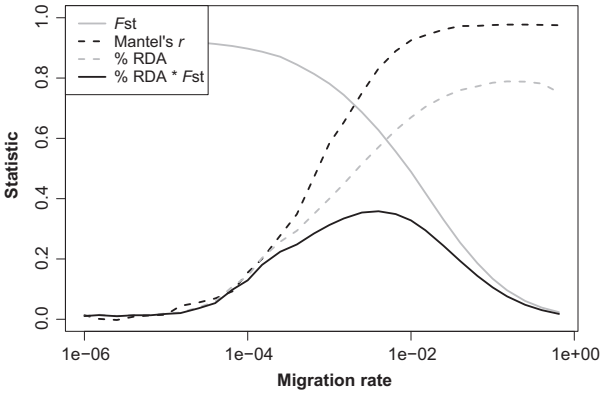

Fig. 2 F_{ST}, Mantel's r and the percentage of constrained variation of an redundancy analysis (RDA) as functions of the migration rate under a one-dimensional migration model. Note that high values of Mantel's r do not indicate limited dispersal and therefore Mantel's r is unsuitable for quantifying the importance of isolation by distance. Results are averaged over 100 repeat simulations of a linear metapopulation consisting of 20 populations. See the Supporting information for more information regarding the simulations and the analysis of the simulated data and for the used R script.

What to do. Rather than Mantel's r, use a method that properly decomposes genetic variance. One of the best methods is presented by a redundancy analysis (RDA), performed on the population allele frequencies (Orsini *et al.* 2012; Wang *et al.* 2012). RDA is a combination of principal component analysis (PCA) and multiple regression that can be used to assess the influence of a matrix of independent variables (here the geographical coordinates, or polynomial combinations thereof) on a matrix of dependent variables (here the allele frequencies). The percentage of variation explained by all constrained RDA axes then indicates the spatial variance component (Borcard *et al.* 1992). In a population genetic context, this then indicates the spatial component of the among-population variation. However, we are mostly interested in the total genetic variation. The spatial component of the total genetic variation can be obtained by making use of the link between PCA and F_{ST}. It has previously been shown that the total variation of a PCA on allele frequencies is equivalent to F_{ST} (Goudet 1999; McVean 2009). In an RDA, we can therefore multiply the percentage of constrained variation by the overall value of F_{ST} to obtain the percentage of the total genetic variation that is explained by the spatial variables (for calculations, see the R script in the Supporting information). For the simulated data, we see that this approach does yield the expected pattern: the greatest fraction of spatially bound variation is observed at intermediate migration rates (Fig. 2). An RDA is very flexible and cannot only be used with spatial variables as explanatory vari-

ables, but also with environmental or other types of data. By specifying variance components (or even individual RDA axes) in terms of F_{ST}, RDA can be used to assess the relative importance of IBA and IBD on the total genetic variance rather than only the among-population variance.

Forgetting that only a small portion of the genome will be associated with climate

In a large genomewide association study (GWAS), Fournier-Level *et al.* (2011) tested for climatic adaptation in *Arabidopsis thaliana* using a set of 213 248 SNPs. After applying a set of criteria to avoid false positives, they presented a list of only four SNPs that were associated with fitness in field conditions. At only 0.002% of all genotyped SNPs, this number is low, but not surprisingly so. After all, even small genomes such as that of *A. thaliana* are pretty big, most of the genome is nonfunctional, and the functional regions mostly have little to do with climatic adaptation. In addition, there is the problem of missing heritability (Yang *et al.* 2010): even studies that genotype thousands of individuals at large numbers of SNPs generally uncover only a small proportion of the heritable variation in a trait. For example, the full model of Fournier-Level *et al.* (2011) only explained 9–24% of the variation in survival, which they called 'a substantial amount'. Therefore, the 0.002% of SNPs that they found to be involved in climate adaptation seems to be in the correct order of magnitude.

Most studies that look for GEA do not employ the common garden approach that makes the GWAS study of Fournier-Level *et al.* (2011) so powerful. Such studies do not have the sampling scope, nor the genomic tools that were used in the *Arabidopsis* study. Given that the probability of finding loci under selection is quite low, most GEA tests are unlikely to uncover any genes at all that are directly under selection (Tiffin & Ross-Ibarra 2014). However, this is not what we see in studies that use GEA tests: it is not uncommon for AFLP or RAD tag studies to find that around 5% of the markers are putatively under environmental selection (unsurprisingly, close to the used alpha level to assess significance). Are we then to believe that 5% of randomly sampled variation across the genome, which is largely nonfunctional, is showing adaptation for exactly the climatic variables that the authors are studying? I find that hard to swallow as genomewide association studies, which are much more powerful than genome scans, usually only find a handful of genes.

Often, there are alternative explanations that are more parsimonious than assuming selection. First, there are ample ways in which a GEA analysis may have been biased. As the actual occurrence of climate-linked loci in the genome will be orders of magnitude lower than the standard alpha level of 0.05, a high rate of false positives is expected. Assume, for example, that the analysis of Fournier-Level *et al.* (2011) has been extremely conservative and that they picked out only 1% of the actual climate-related SNPs in their data set. In such a case, the actual frequency of true positives would still be only 0.2%. In contrast, the probability of getting a false positive is 5%: 25 times as high. Reducing the type I error through adjustment of the alpha level or correction for multiple testing may help to a degree, but at the cost of inflating the type II errors. Additional false positives may arise from the presence of hierarchical population structure (Excoffier *et al.* 2009) and IBD (Meirmans 2012a). Furthermore, many other factors, such as incorrect sampling (Lotterhos & Whitlock 2015), may lead to similar biases, although there is surprisingly little theoretically exploration of this. Second, it is possible that associations among environmental variables and genetic variation arise without invoking environmental adaptation. Under the coupling hypothesis (Bierne *et al.* 2011), genomic incompatibilities may cause tension zones, whose location can move due to stochasticity. Over time, such tension zones have the tendency to become coincident with ecological barriers, leading to a statistical association between the ecological factors and the involved genomic regions. This process is much more likely to lead to the detection of GEA as they concern much larger genomic regions (Bierne *et al.* 2011) and are therefore much more likely to pop up when only a limited number of loci are sampled.

What to do. Use a method that avoids bias stemming from confounding factors (e.g. Excoffier *et al.* 2009; Frichot *et al.* 2013; Günther & Coop 2013). The use of multiple methods concurrently may help you to select the loci that show the most robust pattern; be careful, however, that different methods may be sensitive to similar biases. The power of several of these methods has been shown to largely depend on the sampling strategy (Lotterhos & Whitlock 2015), so a good sampling strategy will help to uncover the actual genes under selection and reduce false positives. However, for the reasons outlined above, it is nearly impossible to completely avoid false positives. Perhaps it is more fruitful to change the way we approach these questions. From an ecological viewpoint, it is often of more interest to study the processes that lead to the adaptation to the local climate, than to pinpoint the actual loci. This is especially the case when using anonymous markers such as AFLPs or RAD tags or when studying species for which very little genome annotation is available. Instead of asking 'Which loci are involved in climate adaptation?' we should rather ask 'Which environmental variables are most important in explaining the distribution of genetic variation?'. Analyses of GEA can therefore be better answered by testing the overall effects of climatic variables, rather than carrying out a separate test for every gene. This immediately solves the problem of reduced power as researchers often have good a priori hypotheses of which climatic variables may affect their species and there are usually much less climatic variables than there are genes, leading to less testing and less need for correction. As in the estimation of the spatial component of the genetic variation, an RDA can be used very effectively to estimate the genetic variance components associated with

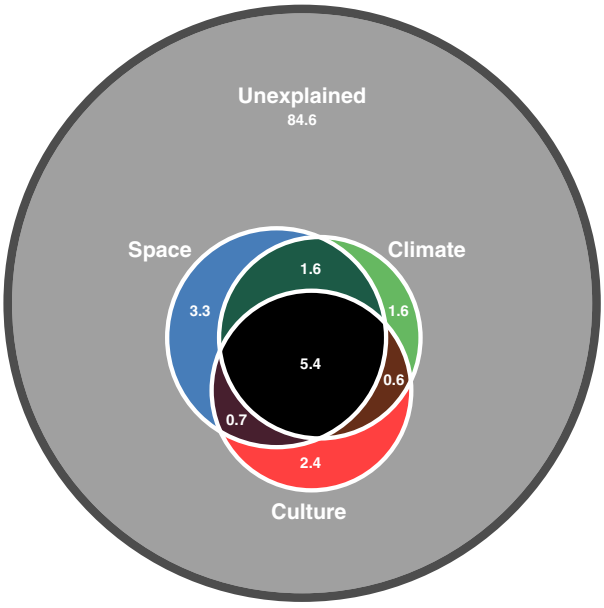

Fig. 3 Using redundancy analysis (RDA) to decompose the among-population variation of 120 African populations of humans (data from Tishkoff *et al.* 2009) into spatial, climatic and cultural components. The grey circle illustrates the total among-population variation, the three coloured circles the variation explained by the three explanatory matrices and their overlap the fraction of shared variation. The high fraction of unexplained variation is probably due to genetic drift within populations, which does not lead to any statistical association with the explanatory variables. Given an overall F_{ST} of 0.015, the 16.2% of explained variation is equivalent to an F_{ST} of 0.0024, illustrating (as expected) that very little of the total genetic variation in humans is associated with any of the explanatory variables. See the Supporting information for more information regarding the data analysis and for the used R script.

different sets of climatic variables and their degree of overlap (Orsini *et al.* 2012; Wang *et al.* 2012). Figure 3 illustrates this approach using data from humans from 120 African populations (Tishkoff *et al.* 2009). It shows that culture—language and subsistence mode—explains slightly more of the genetic variation among populations (3.0%) than climate (2.2%), after correction for the influence of spatial processes. However, most of the explained variance is shared among the three sets of explanatory variables.

Conclusions

When looking at the above advice on how to avoid those seven common mistakes, there is one main message that stands out: for all genotyping experiments, researchers need to consider the experimental design and appropriate analyses prior to execution. This means that there should be a well-designed sampling strategy that is balanced, prevents pseudoreplication and accounts for possible confounding factors. Furthermore, as in any experiment, proper randomization should be employed during genotyping to prevent bias.

When analysing the data, it should be clear what part of the analysis consist of testing a priori hypotheses, and which parts are more of an exploratory nature. Exploratory analyses are an important part of population genetics, as every species has an idiosyncratic pattern in the distribution of its genetic variation. This is due to the uniqueness of its demographic history and the stochasticity inherent in the genealogical processes. Therefore, it is often hard to formulate specific a priori hypotheses, as those would require detailed information about these unknown demographic processes. When testing a priori hypotheses, these should be specified clearly, based on explicit ecological or evolutionary theories or on independent data, not on results of exploratory analyses of the same data.

When interpreting the results, it is important to focus more on biological relevance than on statistical significance. That does not mean that significance is unimportant; results that have a straightforward interpretation but are not significant should not be considered. On the other hand, one should not be blinded by results that are strongly significant. In the genomics era, with thousands upon thousands of loci, strong significance is easily obtained even for biologically marginal processes. For example, I have shown here that a high value of Mantel's r can be strongly significant but may be biologically irrelevant, as it explains only a very small proportion of the total genetic variation and says little about dispersal limitations. This reflects the results of Puebla *et al.* (2012), who noted that very limited dispersal may actually result in nonsignificant Mantel tests. Furthermore, researchers should be more aware of the limitations of their analyses and the resulting uncertainties in the outcome. As estimating the correct number of clusters k is difficult, it is wise to consider other values as well that have strong support. When tests for genotype-by-environment associations are known to lead to high amounts of false positives (Lotterhos & Whitlock 2014), one should not accept the results at face value, especially when an unrealistically high number of selected loci are returned. The inclusion of ever-increasing numbers of loci in genotyping studies also means that we will have to step away from methods that were designed for single loci and move towards analyses that take linkage between thousands of loci into account (Lawson *et al.* 2012).

All in all, modern genomic tools present an enormous opportunity for the field of molecular ecology, potentially enabling deeper insights into the relationships between genes and ecology than has hitherto been possible. This study has shown, however, that the analysis and interpretation of modern genomic data sets require greater care than they have often received so far. Statistics can be powerful tools for scientific inferences, but, as with any tool, they should be applied correctly and responsibly.

Acknowledgements

I would like to thank many colleagues that have over the years discussed these topics with me. Stephanie Meirmans, three anonymous reviewers and the editor, Nolan Kane, gave valuable comments on an earlier version of the manuscript.

References

Audzijonyte A, Vrijenhoek R (2010) When gaps really are gaps: statistical phylogeography of hydrothermal vent invertebrates. *Evolution*, **64**, 2369–2384.

Betto-Colliard C, Sermier R, Litvinchuk S, Perrin N, Stöck M (2015) Origin and genome evolution of polyploid green toads in Central Asia: evidence from microsatellite markers. *Heredity*, **114**, 300–308.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial-DNA restriction data. *Genetics*, **131**, 479–491.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Francois O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Goudet J (1999) PCAGEN: principal components analysis of gene frequency data (version 1.2). Available from http://www2.unil.ch/popgen/softwares/pcagen.htm.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, **7**, 1–32.

Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, **13**, 110–122.

Hurlbert S (1990) Spatial distribution of the montane unicorn. *Oikos*, **58**, 257–271.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

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

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLOS Genetics*, **8**, e1002453.

Legendre P, Fortin MJ (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, **10**, 831–844.

Legendre P, Legendre L (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Lotterhos KE (2012) Nonsignificant isolation by distance implies limited dispersal. *Molecular Ecology*, **21**, 5637–5639.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

McVean G (2009) A genealogical interpretation of principal components analysis. *PLOS Genetics*, **5**, e1000686.

Meirmans PG (2011) MARLIN, software to create, run, and analyse spatially realistic simulations. *Molecular Ecology Resources*, **11**, 146–150.

Meirmans PG (2012a) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.

Meirmans PG (2012b) AMOVA-based clustering of population genetic data. *Journal of Heredity*, **103**, 744–750.

Milligan G, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Neuenschwander S, Hospital F, Guillaume F, Goudet J (2008) QuantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics*, **24**, 1552–1553.

Orsini L, Mergeay J, Vanoverbeke J, De Meester L (2012) The role of selection in driving landscape genomic structure of the waterflea *Daphnia magna*. *Molecular Ecology*, **22**, 583–601.

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puebla O, Bermingham E, McMillan WO (2012) On the spatial scale of dispersal in coral reef fishes. *Molecular Ecology*, **21**, 5675–5688.

Raufaste N, Rousset F (2001) Are partial Mantel tests adequate? *Evolution*, **55**, 1703–1705.

Shafer ABA, Wolf JBW (2013) Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecology Letters*, **16**, 940–950.

Taberlet P, Fumagalli L, Wust-Saucy A, Cosson J (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

Tishkoff SA, Reed FA, Friedlaender FR *et al.* (2009) The genetic structure and history of Africans and African Americans. *Science*, **324**, 1035–1044.

Van Hengstum T, Lachmuth S, Oostermeijer JGB *et al.* (2012) Human-induced hybridization among congeneric endemic plants on Tenerife, Canary Islands. *Plant Systematics and Evolution*, **298**, 1119–1131.

Vrijenhoek RC, Waples RS (2012) Popular misconceptions. *Molecular Ecology*, **21**, 4155–4156.

Wang IJ, Glor RE, Losos JB (2012) Quantifying the roles of ecology and geography in spatial genetic divergence *Ecology Letters*, **16**, 175–182.

Waples RS (2014) Testing for Hardy–Weinberg proportions: have we lost the plot? *Journal of Heredity*, **106**, 1–19.
Yang J, Benyamin B, Mcevoy BP *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

P.G.M. observed the errors, performed the simulations, analysed the data, took care of the unicorns, and wrote the manuscript.

doi: 10.1111/mec.13243

Data accessibility

Simulated genotype data: uploaded to Dryad http://dx.doi.org/10.5061/dryad.kp905.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1. Simulations and analyses.