Methods in Ecology and Evolution

Methods in Ecology and Evolution 2017, **8,** 717–727 doi: 10.1111/2041-210X.12774

HOW TO MEASURE NATURAL SELECTION

Composite measures of selection can improve the signal-to-noise ratio in genome scans

Katie E. Lotterhos[*,1], **Daren C. Card**[2], **Sara M. Schaal**[1], **Liuyang Wang**[3], **Caitlin Collins**[4] **and Bob Verity**[4]

[1]Northeastern University Marine Science Center, 430 Nahant Rd, Nahant, MA 01908, USA; [2]Department of Biology, University of Texas at Arlington, 501 S. Nedderman Drive, Arlington, TX 76019, USA; [3]Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, Durham, NC 27710, USA; and [4]Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College London, London SW7 2AZ, UK

Summary

1. The growing wealth of genomic data is yielding new insights into the genetic basis of adaptation, but it also presents the challenge of extracting the relevant signal from multi-dimensional datasets. Different statistical approaches vary in their power to detect selection depending on the demographic history, type of selection, genetic architecture and experimental design.

2. Here, we develop and evaluate new approaches for combining results from multiple tests, including multivariate distance measures and methods for combining *P*-values. We evaluate these methods on (i) simulated landscape genetic data analysed for differentiation outliers and genetic-environment associations and (ii) empirical genomic data analysed for selective sweeps within dog breeds for loci known to be selected for during domestication. We also introduce and evaluate how robust statistical algorithms can be used for parameter estimation in statistical genomics.

3. On the simulated data, many of the composite measures performed well and had decreased variation in outcomes across many sampling designs. On the empirical dataset, methods based on combining *P*-values generally performed better with clearer signals of selection, higher significance of the signal, and in closer proximity to the known selected locus. Although robust algorithms could identify neutral loci in our simulations, they did not universally improve power to detect selection. Overall, a composite statistic that measured a robust multivariate distance from rank-based *P*-values performed the best.

4. We found that composite measures of selection could improve the signal of selection in many cases, but they were not a panacea and their power is limited by the power of the univariate statistics they summarize. Since genome scans are widely used, improving inference for prioritizing candidate genes may be beneficial to medicine, agriculture, and breeding. Our results also have application to outlier detection in high-dimensional datasets and to combining results in meta-analyses in many disciplines. The compound measures we evaluate are implemented in the R package MINOTAUR.

Key-words: composite signals of selection, genome-wide associations, Mahalanobis distance, meta-analysis, minimum covariance determinant

Introduction

The rapid improvement of high-throughput sequencing technologies has stimulated studies that examine the genomic basis of adaptation and of phenotypic traits. This progress has been paralleled by the development of new genome scan methods aimed at detecting selection. Genome scans have implicated many genomic variants with effects on adaptive traits in plants (e.g. Savolainen, Lascoux & Merila 2013), animals (e.g. Hoekstra *et al.* 2006; Barrett, Rogers & Schluter 2008), and humans (e.g. Hindorff *et al.* 2009). All genome scans are based on the premise that loci affected by selection will be ‘outliers’ relative to the genome-wide distribution. The loci uncovered by genome scans, however, generally explain a small portion of phenotypic variation (Yang *et al.* 2010; Brachi, Morris & Borevitz 2011). Thus far, the field has been unable to fully characterize the genetic basis of adaptive traits, hampering our ability to understand adaptation.

The major limitation of genome scans is that different methods return inconsistent results (Lotterhos & Whitlock 2015; Schlamp *et al.* 2016; Vatsiou, Bazin & Gaggiotti 2016). Most genome scan methods fall into a univariate framework, in which outliers are identified as a function of one statistic

*Correspondence author. E-mail: k.lotterhos@neu.edu

(Hoban *et al.* 2016; Schlamp *et al.* 2016; Vatsiou, Bazin & Gaggiotti 2016). Variation in results occurs because of differential sensitivity to sampling design (De Mita *et al.* 2013; Lotterhos & Whitlock 2015), to details of the selective sweep (e.g. Schlamp *et al.* 2016; Vatsiou, Bazin & Gaggiotti 2016), and to the demographic history (De Villemereuil *et al.* 2014; Lotterhos & Whitlock 2015; Luu, Bazin & Blum 2016; Schlamp *et al.* 2016). Thus, the complex and largely unknown evolutionary histories of most species make it unlikely that a single statistic can fully capture the genomic signal of interest in the majority of cases (Verity & Nichols 2014; Vatsiou, Bazin & Gaggiotti 2016).

Another limitation of current genome scan methods is that parameter estimates used to control inflation of the test statistic may be biased by genomic regions that are affected by selection (e.g. 'outliers' in the data). For example, background selection on neutral loci linked to deleterious alleles can bias demographic inference (Ewing & Jensen 2016). If the species has widespread linkage disequilibrium, it is still unclear how non-independence among loci, especially in the presence of selection, can change statistical outcomes. Few methods have been developed with the goal of reducing the effects of outliers on parameter estimation (but see Whitlock & Lotterhos 2015).

Given that various methods have different strengths and weaknesses and may be differentially susceptible to outliers, it is difficult to decide how to prioritize candidates for further investigations (François *et al.* 2015). A common approach is to prioritize candidate loci that are outliers in all of the univariate methods (i.e. take the overlap among all univariate statistics), but this approach may miss loci affected by weak selection (Lotterhos & Whitlock 2015). Recently, several composite measures have been proposed based on combining *P*-values and these measures generally perform better than univariate single statistics, but evaluation of their performance has been limited to relatively simple scenarios (Grossman *et al.* 2010; Evangelou & Ioannidis 2013; Utsunomiya *et al.* 2013; Randhawa *et al.* 2014, 2015; François *et al.* 2015; Ma *et al.* 2015).

Here, we describe new approaches for combining signals across test statistics in multivariate space and compare them to other composite measures of selection. We also explore the potential of so-called 'robust' algorithms, which aim to identify a subset of data points that are not outliers in multivariate space, as a way of improving power and reliability. We compare the performance of composite measures to univariate methods for detecting selection on two sets of data (simulations and empirical data), with and without the robust algorithms for parameter estimation. The benefit of the simulated dataset is that the neutral and selected loci are known, although loci are unlinked and so methods based on haplotype structure cannot be evaluated. The benefit of the empirical dataset is that methods can be evaluated on their ability to detect loci known to be selected for during domestication against realistic patterns of linkage disequilibrium and haplotype structure, although neutral loci are unknown.

Materials and methods

COMPOSITE MEASURES BASED ON MULTIVARIATE DISTANCES

Multivariate-distance measures identify points that are distant from the main mass of points. These composite measures are directionless (meaning the idea of upper and lower tails does not apply) and so they are most appropriate for identifying points that deviate from the mass of points by a large amount in any direction. In multivariate statistics, the multivariate mean is known as the 'location' (denoted $\bar{x}$ in Fig. S1, Supporting Information) of the data and the covariance among univariate statistics is called the 'scatter' (matrix denoted *S* in Fig. S1).

Multivariate distance measures: Mahalanobis distance (M_d), harmonic mean distance (H_d), nearest neighbour distance (N_d)

The Mahalanobis distance (M_d; Mahalanobis 1936) is a widely used distance measure relating a point to the multivariate location (Fig. S1a). The M_d differs from the ordinary Euclidean distance due to the correction for covariance among observations. Because the Mahalanobis distance assumes the data can be described by a multivariate ellipsoid, it will tend to perform poorly when observations have a nonparametric or multi-modal distribution (see Verity *et al.* 2017 for visualization).

Previously we developed other multivariate distance metrics that are closely related to M_d in that they correct for covariance among variables, but unlike M_d they relax the assumption that data must be parametric (Fig. S1, Verity *et al.* 2017). Here, we evaluate two of these measures. The first measure is the harmonic mean distance (H_d, Fig. S1b) from a focal point to all other points in the dataset, which tends to perform similarly to M_d (Verity *et al.* 2017). The second measure is the nearest neighbour distance (N_d, Fig. S1d), which measures the shortest distance from the focal point to any other data point. N_d behaves differently than M_d and H_d because it is much more sensitive to the density of local points around the focal point (Verity *et al.* 2017). A third measure of deviance based on kernel density was not used because it was too slow to calculate (Fig. S1c).

Hierarchical clustering (Hclust)

Clustering algorithms identify outliers as data points that do not cluster with other data points. Among available clustering algorithms, a preliminary evaluation found that Ward's minimum variance method had the highest performance (results not shown). Agglomeration by Ward's method joins points into clusters recursively, by choosing the merge that causes the minimum increase in total within-cluster variance at successive stages in the clustering process (Ward 1963; Murtagh & Legendre 2014). The distance between two clusters is calculated from the ANOVA sum of squares, i.e. the sum of the squared pairwise distances across all points in the two clusters. We implemented *Hclust* using the function outliers.ranking() in the R package DMWR 0.4.1 (Togo 2010).

COMPOSITE MEASURES BASED ON COMBINING *P*-VALUES

Methods based on *P*-values have the advantage of quantifying outliers using the familiar concepts of probability and statistical significance.

They also exist on an absolute scale, making comparisons between methods straightforward. The composite measures we employ are based on *P*-values created by ranking the data. These rank-based *P*-values are not *P*-values in the classical statistical sense, but reflect quantile values from the empirical distribution of the test statistic. The drawback of the ranking approach is that the *P*-value distribution is uniform and loci with a strong signal of selection may be less significant when ranked compared to a well-calibrated statistical test, while on the other hand the rank transformation may be beneficial if the statistical test is not well calibrated (shown conceptually in Fig. S2).

Composite selection signals (CSS)

This approach employs Stouffer's method for combining *P*-values, which assumes independence among the statistics being summarized (Stouffer *et al.* 1949; Whitlock 2005). The CSS statistic is calculated from univariate measures as follows: (i) convert univariate statistics to fractional ranks between $1/(n + 1)$ and $n/(n + 1)$, where n is the number of observations, (ii) convert fractional ranks to *Z*-values using the inverse normal cumulative distribution function, (iii) take the mean *Z*-score and convert to a *P*-value using the normal $N(0, m^{-1})$ distribution, where m is the number of univariate statistics, (iv) the CSS statistic is defined as $-\log_{10}$ of the *P*-value (Randhawa *et al.* 2014, 2015). Note that this method does not account for covariance among signals in univariate statistics, nor directionality in the signal.

De-correlated composite of multiple signals (DCMS)

The DCMS is similar to CSS, but does not assume independence among univariate statistics. DCMS is based on the sum of $\log_{10}((1-p)/p)$ over all univariate statistics divided by a weighting vector for each locus (Ma *et al.* 2015). The weights are determined by the genome-wide correlation between all pairs of univariate statistics, such that highly correlated statistics contribute less to the calculation. For example, if two statistics are perfectly correlated and a third statistic is uncorrelated with the first two, the respective weights will be (½, ½, 1). For DCMS, we transformed raw statistics into *P*-values via fractional ranks between $1/(n+1)$ and $n/(n+1)$, using one-tailed or two-tailed rankings as required.

Mahalanobis distance based on negative-log rank-based P-values (Md-rank-*P*)

Md-rank-*P* is computed as the Mahalanobis distance on the negative $\log_{10}$ on the transformation of raw statistics into rank *P*-values (as described for DCMS) from a multivariate location of 0 (a non-significant value) in all dimensions. Md-rank-*P* differs from CSS and DCMS because it measures the distance of an observation from a universally non-significant value in multivariate space, and unlike Mahalanobis distance is it based on *P*-values and not the test statistic or other effect size.

DEFAULT VS. ROBUST APPROACH

Many of the methods that we evaluate take into account the covariance structure among test statistics. The presence of outliers, however, may bias estimation of this covariance matrix. For instance, a small proportion of loci under very strong selection with strong signals across multiple test statistics would increase the overall covariance and bias the multivariate mean, while there would be no correlation (i.e. no bias) if only neutral regions were considered (e.g. Grossman *et al.* 2010).

We evaluated the minimum covariance determinant (MCD) algorithm for identifying 'robust' points (i.e. points that are not outliers in any one dimension) in multivariate genomic data and used these points in the calculation when relevant. The MCD identifies robust points as the set of points that minimize the volume of an ellipsoid surrounding the data in multidimensional space (mathematically the ellipse is described by the determinant on the covariance matrix, Rousseeuw & Driessen 1999). The MCD requires the user to input the proportion of the dataset that will be used for the algorithm with the requirement that the proportion is between 0·50 and 1. We implemented the MCD using the proportion 0·75, as recommended by Rousseeuw & Driessen (1999; preliminary analyses revealed that the results were not sensitive to this proportion), with the function CovNAMcd in the R package RRCOVNA (Todorov & Filzmoser 2009). In preliminary analyses, the MCD outperformed a different robust method called the projection congruent subset (Schmitt, Oellerer & Vakili 2014; Vakili & Schmitt 2014, results not shown).

LANDSCAPE SIMULATIONS

To test the power of multidimensional outliers for genome scans, we applied them to published simulated datasets sampled from landscape simulations (Lotterhos & Whitlock 2014, 2015). Briefly, a landscape simulator was used to simulate haploid loci in four demographic histories: island model (IM), isolation by distance (IBD), expansion from one refuge (1R), and expansion from two refugia (2R). Selected loci were simulated under varying strengths of selection (s_L) to a heterogeneous latitudinal cline in an environmental variable that affected fitness. Datasets consisted of 9900 neutral and 100 selected loci in individuals randomly sampled from the landscape in six different ways. The distributions of the strengths of selection were varied to make the response to selection more equivalent across datasets. For IBD, the dataset included four strengths of selection in each demography at the following percentages: $s_L = 0{\cdot}001$ (40% of the loci), $s_L = 0{\cdot}005$ (30%), $s_L = 0{\cdot}01$ (20%) and $s_L = 0{\cdot}1$ (10%). For IM, 1R and 2R, the dataset included three strengths of selection at the following percentages: $s_L = 0{\cdot}005$ (50%), $s_L = 0{\cdot}01$ (33%) and $s_L = 0{\cdot}1$ (17%). For details see Lotterhos & Whitlock (2015).

Lotterhos & Whitlock (2015) used these datasets to perform univariate genome scans in the programs BAYENV2 (Günther & Coop 2013) and LFMM (Frichot *et al.* 2013; Frichot & François 2015). These programs were designed for landscape genetic datasets with environmental data. A total of four univariate statistics from these two programs were combined to produce composite measures: (i) log-Bayes Factor (BF, association between allele frequency and the environment, BAYENV2), (ii) Spearman's ρ (association between allele frequency and the environment, BAYENV2), (iii) X^TX (genetic differentiation among populations, BAYENV2), and (iv) *Z*-score (association between genotype and the environment in LFMM). The power of these four univariate statistics varied with sampling design and demographic history (Lotterhos & Whitlock 2015). Note that given recent improvement to the algorithm in BAYENV2 has been implemented in the program BAYPASS (Gautier 2015), we compared output from these programs and found only slight differences in performance (see Dryad Repository). Here, we use the BAYENV2 results so that this manuscript can be directly compared to Lotterhos & Whitlock (2015).

Comparison of robust points and neutral loci

For each dataset, we evaluated the mean absolute difference between the actual multivariate location or scatter of the neutral points, and an

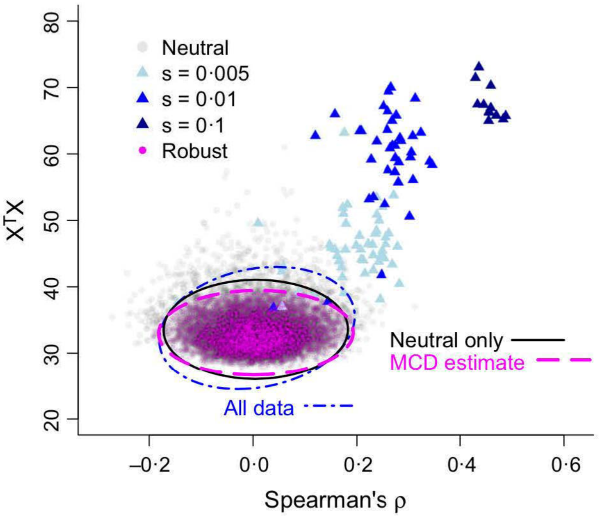

Fig. 1. For a single dataset simulated under the two refugia model, comparison of 95% confidence intervals on the two dimensional ellipse given by the determinant on the covariance matrix between these two variables calculated from: all the data (dashed-dotted blue line), neutral loci only (solid black line), and the minimum covariance determinant (MCD) estimate (dashed magenta line). Loci simulated under selection are triangles, while loci identified as robust by the MCD are highlighted in magenta.

estimate of the location and scatter based on either all the data or the MCD subset of points. We also evaluated whether the robust points contained fewer loci simulated under selection than would have been expected by chance.

Calculation of empirical power

True positive rate or power is the proportion of loci simulated under selection that are identified as candidates. The empirical power is calculated as follows: (i) use all known neutral loci to generate a null distribution; (ii) for each locus calculate an empirical *P*-value based on its cumulative frequency in this null distribution (for details see Lotterhos & Whitlock 2015). Empirical power minimizes the false positive rate and makes it equal across all comparisons, such that test statistics are compared on common grounds in their ability to separate signals of neutral vs. selected loci. To control for false discovery rate, we converted *P*-values to *q*-values using the Benjamini & Hochberg algorithm, and retained loci with a *q*-value < 0·05 as candidates (a *q*-value of 0·05 has a desired rate of 5 false positives out of 100 positive hits; Benjamini & Hotchberg 1995). Using the neutral loci as an empirical distribution, our *P*-values are perfectly calibrated and therefore meet the assumptions for transformation into *q*-values. Empirical power also has the advantage of being inversely related to false discovery rate. Empirical power is explained in more detail in Fig. S3.

EMPIRICAL DATASET: DOG BREEDS

We employed a recently published study that evaluated scans for selective sweeps in 25 dog breed genomes. Schlamp *et al.* (2016) evaluated eight selective sweep statistics and found variation in their ability to detect 12 quantitative trait loci (QTL) with effects on dog phenotypes known to be under positive selection during domestication. These statistics measure within-breed haplotype structure and sequence diversity, and thus represent an application of composite measures in a different situation from the simulations (and thus the univariate statistics are not cross-compatible).

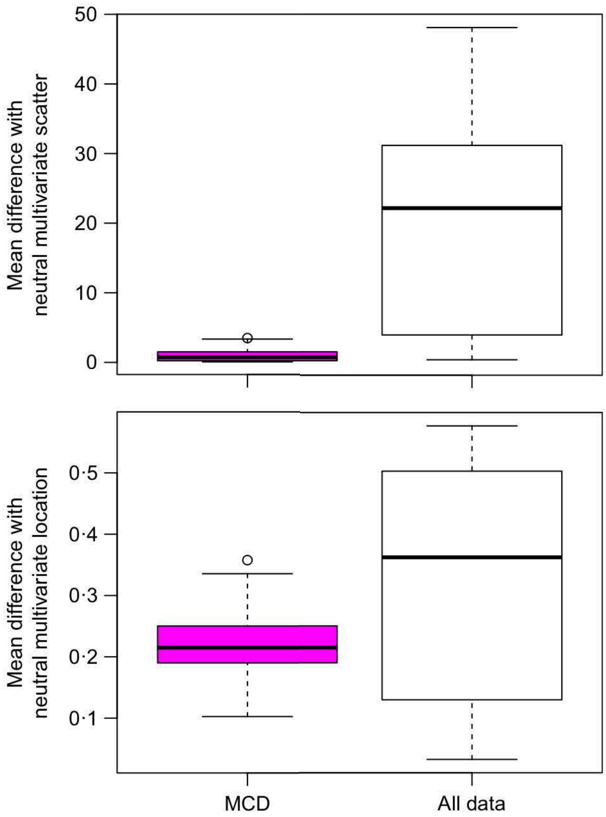

Fig. 2. Mean difference between the neutral estimate and the other estimates of multivariate scatter (top) and multivariate location (bottom), summarized over all demographies and sampling designs. 'MCD' refers to the minimum covariance determinant estimate (using only the robust points), and 'All' refers to all the data (including selected loci). Lower values indicate that the given estimate is closer to the true value. MCD, minimum covariance determinant.

For each of the 25 breeds, we evaluated composite signals based on four of the single measures applied within breeds: iHS (Voight *et al.* 2006), H12 (Garud *et al.* 2015), Tajima's *D* (Tajima 1989a), and nucleotide diversity (π; Nei & Li 1979). iHS is a measure of the extent of haplotype homozygosity, and infers selection by identifying haplotypes that are much larger than expected under neutrality. H12 represents a composite haplotype frequency of the first and second most abundant haplotypes and is designed to detect soft selective sweeps. Tajima's *D* is a variance-standardized measure that may indicate a selective sweep when it is negative. π is the nucleotide diversity and is also expected to have low values under a selective sweep. We excluded four other measures evaluated by Schlamp *et al.* from our analysis: HAPFLK (Fariello *et al.* 2013), because the statistic is calculated across breeds and we were interested in evaluating power within each breed; nSL (Ferrer-Admetlla *et al.* 2014), because the univariate statistic showed no meaningful signal; and the measures H (Messer 2015) and composite likelihood ratio (CLR; calculated, using method of Pavlidis *et al.* 2013; but also see Kim & Stephan 2002 and Nielsen *et al.* 2005), because of minimal overlap with other statistics, resulting in large amounts of missing data. Schlamp *et al.* (2016) compared calculation of these statistics under different window sizes (based on the number of segregating sites) and found that window sizes of 25 or 51 single-nucleotide polymorphisms (SNPs) generally had clearer signals of selection. We therefore used the 51-SNP window sizes to leverage this increased power while deriving smoother results.

Evaluation of statistics

Because neutral loci in this dataset are unknown, we cannot calculate empirical power in the same manner as the simulated data. Instead, we compared the ability of different measures to detect known QTL that had risen to a frequency of at least 50% in each breed (following Schlamp *et al.* 2016, hereafter: focal SNP), based on whether there was a signal in the region surrounding the focal SNP. We quantified the signal of selection in three ways (shown conceptually in Fig. S4):

1. The significance of the signal was measured as the observed quantile of the SNP with the most extreme signal (maximum or minimum, depending on the tail of the test) within a 51-SNP window centred on the focal SNP. The quantile of the extreme SNP was calculated from ranking all SNPs on that chromosome and transformed to a $-\log_{10}$ *P*-value as described in the DCMS section.

2. The clarity of the signal in the window on either side of the extreme SNP was quantified based as a steepness measure (hereafter: steepness, Fig. S4). The slope (m) on either side of the extreme SNP was calculated from a linear model for a 20-SNP window either upstream or downstream of the extreme SNP, and averaged as steepness $= (m_{\mathrm{up}} \times I_{\mathrm{up}} + m_{\mathrm{down}} \times I_{\mathrm{down}})/2$, where I_x is an indicator variable depending on whether the test is in the upper or lower tail of the test statistic (visualized in Fig. S4). This steepness value has the desirable property that it is large and positive if there is a peak near the extreme SNP (indicating a signal of selection), that it is near zero if there is no signal, and it is negative if the concavity of the signal is in the opposite direction (Fig. S4). For the calculation of steepness, for each slope we tested if it was significant at $P < 0{\cdot}01$ and if it was not significant then the slope was assigned a value of 0 (to avoid averaging spuriously large slopes with large standard error). Note that steepness should be interpreted with caution because LD will affect the steepness of the sweep.

3. The distance (in Kb) between the extreme SNP and the focal SNP (Fig. S4).

Results

LANDSCAPE SIMULATIONS

Comparison between robust and neutral estimates

A typical comparison between the MCD estimate of the covariance, the neutral estimate of the covariance, and the estimate using all the data is visualized in Fig. 1 for the 2R model. In this case, the MCD estimates of location and scatter more accurately captured the neutral estimates compared to the estimates that used all the data. Figure 1 also shows how many neutral loci were not identified as robust points by the MCD because of statistical noise.

For all the simulated datasets, the robust estimates of location and scatter from the MCD was generally a more reliable estimate of neutral location and scatter (Fig. 2, 'MCD') than the estimate, using all the data (Fig. 2, 'All').

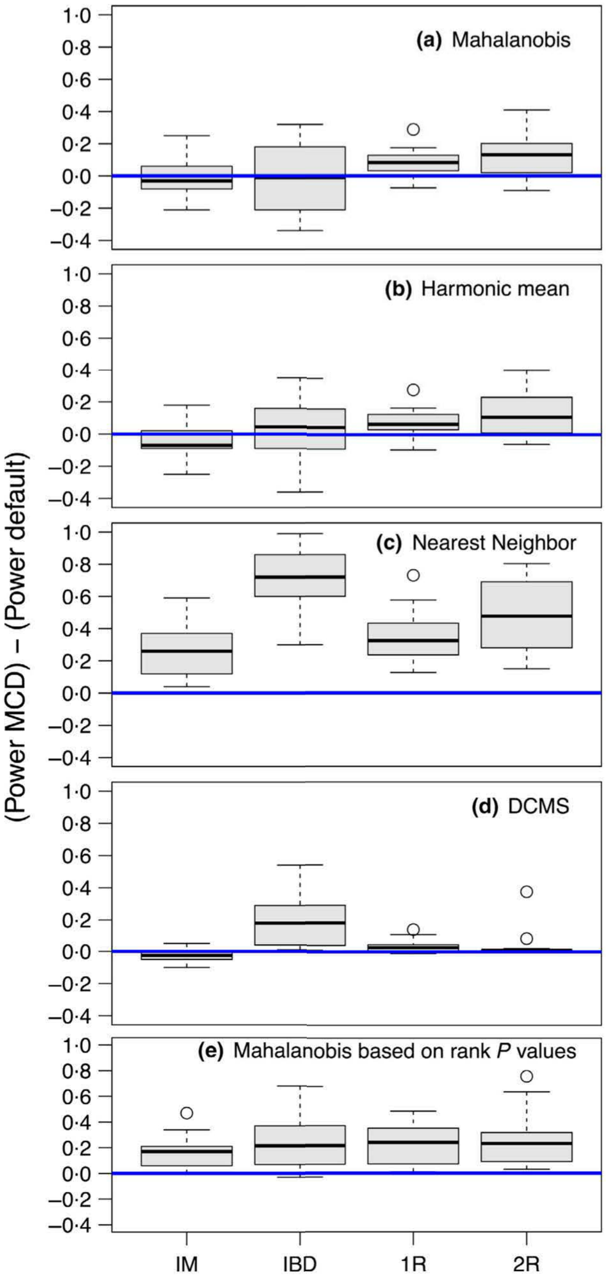

Fig. 3. The difference in empirical power of the statistic to detect selection with the minimum covariance determinant minimum covariance determinant (MCD) minus the default value for the four demographies: island model (IM), isolation by distance (IBD), range expansion from one refuge (1R), range expansion from two refugia (2R). (a) Mahalanobis distance, (b) harmonic mean distance, (c) nearest neighbour distance, (d) de-correlated composite of multiple signals (DCMS, combines rank *P*-values), (e) Mahalanobis distance based on rank *P*-values.

The difference, however, was quite variable across all simulations and sampling designs (Fig. 2). The MCD gave better estimates of neutral location and scatter because it contained a smaller proportion of selected loci than expected by chance (Fig. S5). The probability that the locus was included in the list of MCD robust points was negatively

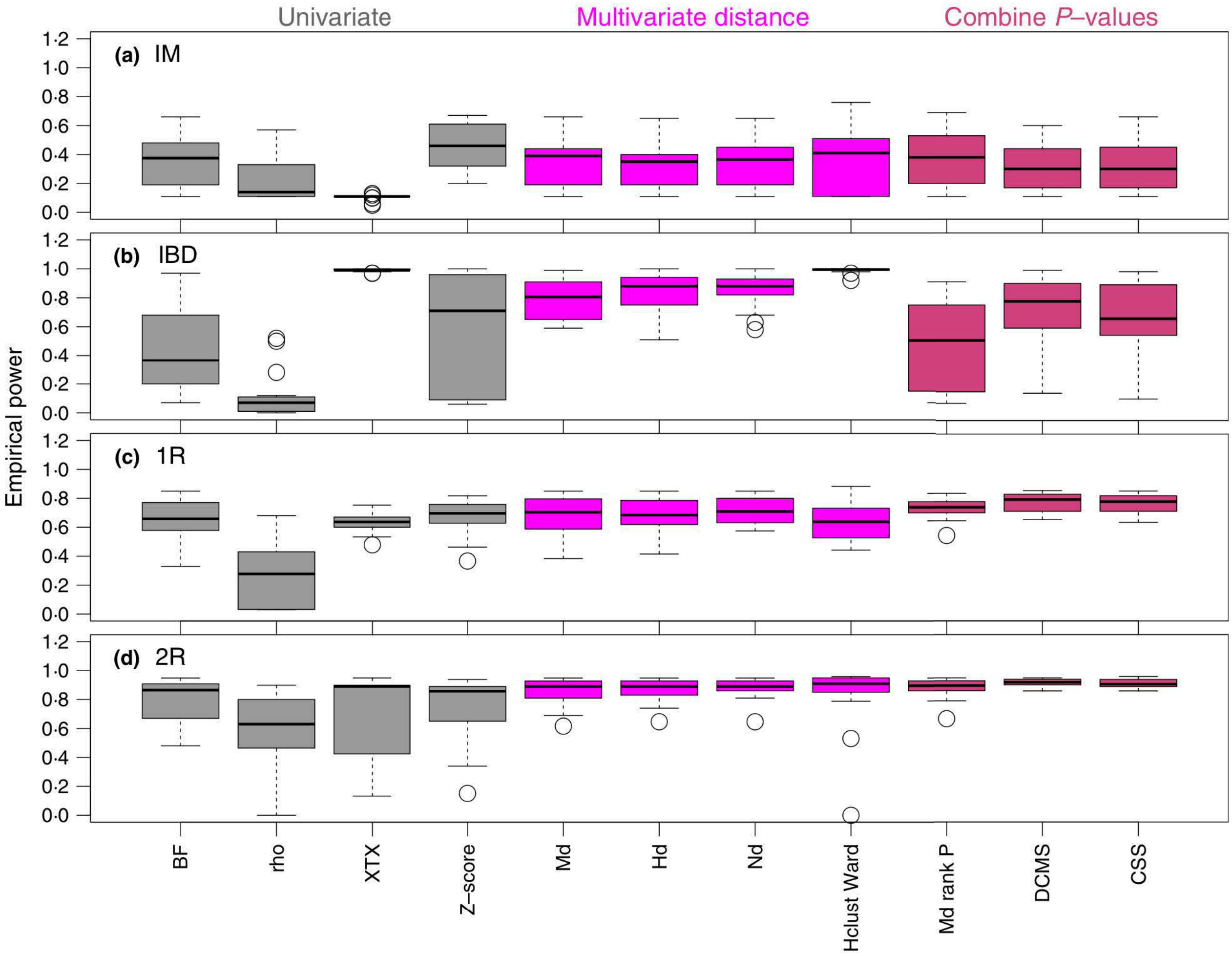

Fig. 4. Comparison of empirical power of the univariate statistics and compound measures evaluated from the demography simulations. Four demographies were evaluated: (a) island model, (b) isolation by distance, (c) range expansion from a single refuge, and (d) range expansion from two refugia. The univariate statistics include Bayes factor (BF) and Spearman's ρ from a genetic-environment association, a measure of genetic differentiation (XTX), and the *Z*-score from the program LFMM. The compound measures based on multivariate distances include: Mahalanobis distance (M_d), harmonic mean distance (H_d), nearest neighbour distance (N_d), and hierarchical clustering (*Hclust Ward*). The compound measures based on combining *P*-values include: decorrelated composite of multiple signals (DCMS), composite signal of selection (CSS), and Mahalanobis distance based on rank *P*-values (Md-rank-*P*). Note that for M_d, H_d, N_d, DCMS and Md-rank-*P* we plot the minimum covariance determinant (MCD) estimate because it either improved or did not affect the overall average power.

correlated with the simulated strength of selection on that locus, and the MCD never identified loci simulated under strong selection as robust (Fig. S5).

Comparison of empirical power with and without a robust approach

Incorporating the MCD algorithm into the calculation could improve power to detect selection, but this effect was not universal (Fig. 3). For Mahalanobis distance and harmonic mean distance, the MCD improved power in the 1R and 2R models (cases with stronger correlations among loci), and had variable effects on power in the IM and IBD models (cases with weaker correlations among loci) (Fig. 3a,b). For nearest neighbour distance, the increase in power was most pronounced (from a 10% to 100% increase, Fig. 3c) because with the MCD the closest neighbour to a locus under selection would be a robust point (i.e. a neutral locus with high probability), whereas without the MCD, the closest neighbour to a locus under selection would be another locus under selection (visualized in Fig. 1). For DCMS, the MCD improved power for the IBD case (Fig. 3d), which is the case with the most variable power in univariate statistics (see next section), and did not substantially change power for the other demographies. For Md-rank-*P*, the MCD always had a moderate improvement in power (5% to 35% increase, Fig. 3e).

Comparison of empirical power for univariate vs. composite measures

For the univariate statistics, power varied greatly among the demographies. For example, X^TX had high empirical power under the IBD model compared to the IM (Fig. 4, grey boxes). Each of the bars plotted in Fig. 4 is summarized over the 6 random population sampling designs, and so the height of the

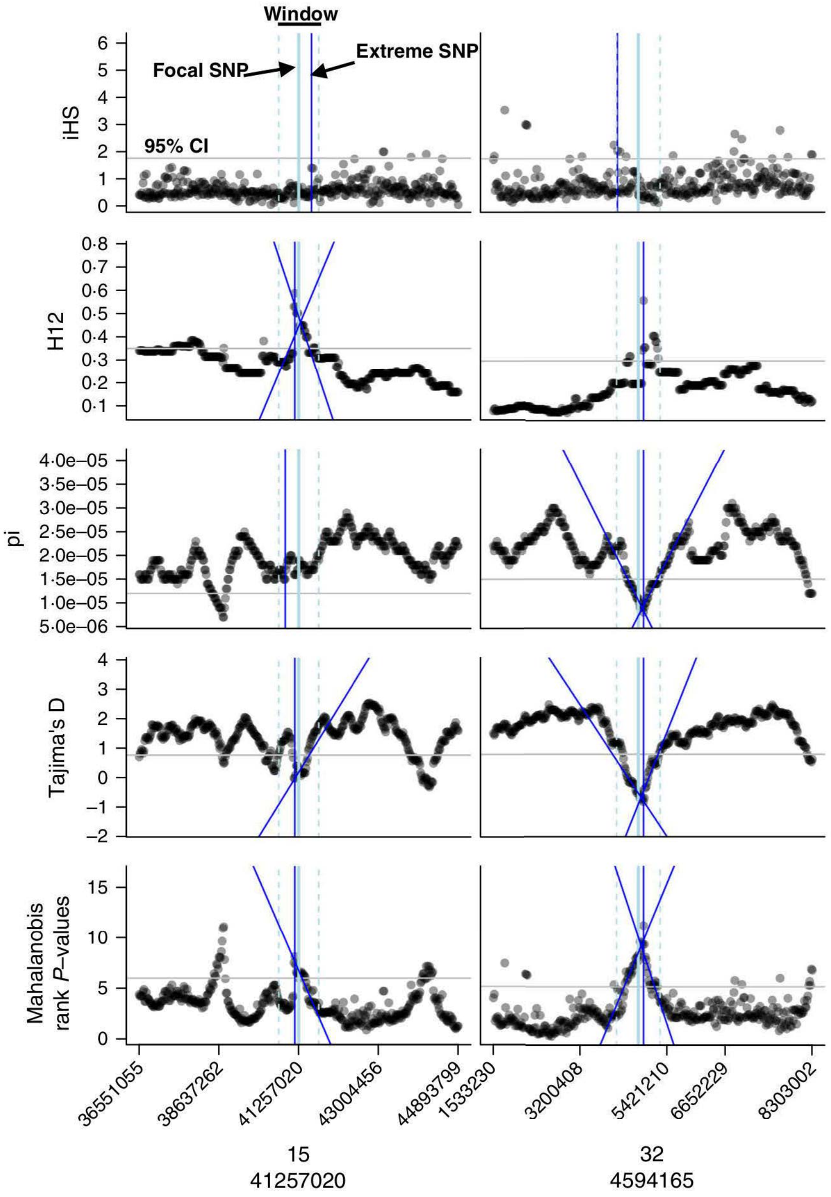

Fig. 5. Examples of selection signals on focal single-nucleotide polymorphisms (SNPs) on chromosomes 15 (left column) and 32 (right column) in the maltese dog breed. The vertical light blue line represents the focal SNP and the vertical light blue dashed lines represent the 51-SNP window centred on the focal SNP. The vertical dark blue line represents the location of the SNP with the most extreme signal in the window, and the slopes leading up to that SNP are shown if they are significant at a *P*-value <0·01 in a general linear model. The horizontal grey line is the appropriate 95% quantile for that statistic on the chromosome (an upper tail for iHS, H12, and Mahalanobis; a lower tail for π and Tajima's *D*).

boxplots illustrates how sampling design can cause variation in outcomes. For instance, *Z*-score and BF had higher variance in results under IBD but lower variance under the 2R model, while X^TX showed the opposite pattern (compare height of grey box plots in Fig. 4b–d).

The power of the compound measures compared to the univariate statistics varied among demographies. For the IM, all compound measures had similar power and variance to the univariate statistics (Fig. 4a). For the IBD model, multivariate distances and hierarchical clustering had higher power and lower variance in power compared to methods that combine *P*-values (DCMS, CSS, Md-rank-*P*) (Fig. 4b). Hclust performed the best in this model by capitalizing on the performance of the X^TX statistic, which was highly informative (in contrast to the IM case). Ward's minimum variance method clustered neutral loci together before merging any of the loci under selection into this 'neutral' cluster, because this kept the total within-cluster variance at a minimum. Methods that combined *P*-values performed poorly in IBD, possibly because of the large variation in the performance of the univariate statistics (Fig. 4b). For the refugia models, all composite measures performed as well as or better than the univariate measures, with substantially lower variance in power across sampling designs and demographies (Fig. 4c,d). The methods that combined *P*-values performed slightly better in the refugia scenarios than the multivariate distances.

EMPIRICAL DATASET: DOG BREEDS

Empirical signals of selection for steepness, significance, and distance of the signal from two causal SNPs in the maltese breed are show in Fig. 5 (also shown conceptually in Fig. S4). Figure 5 also shows how a compound measure can reflect signals from different univariate statistics. In the first case (left column Fig. 5), H12 and Tajima's *D* had a signal at the focal SNP, while in the second case (right column Fig. 5) Tajima's *D*

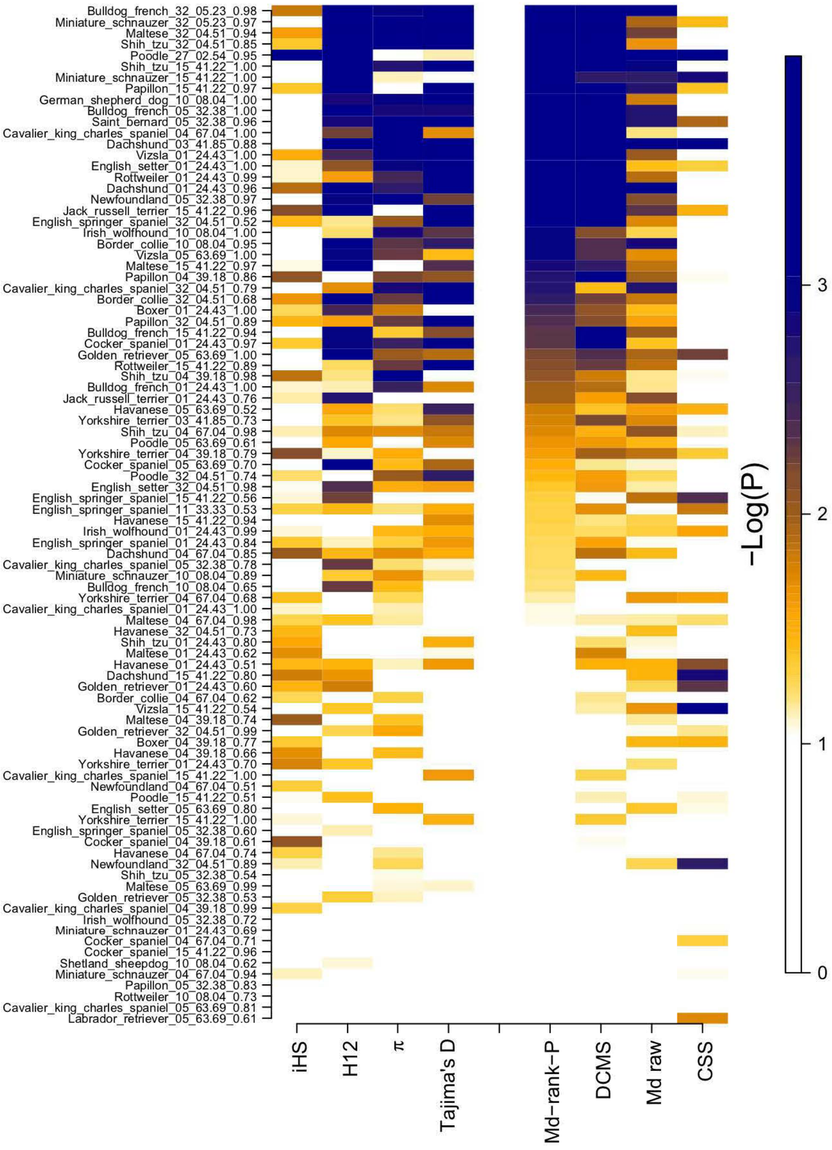

Fig. 6. Heatmap of statistical significance ($-\log_{10}$ fractional rank) evaluated for each individual locus in the empirical dog dataset. The univariate statistics include iHS, H12, π (nucleotide diversity), and Tajima's *D*. The compound measures include: the Mahalanobis distance based on negative-log rank *P*-values (Md-rank-*P*), the decorrelated composite of multiple signals (DCMS), the Mahalanobis distance based on the raw statistics (M_d raw), and the composite signal of selection (CSS). The row labels indicate the dog breed, chromosome, base pair of the focal single-nucleotide polymorphisms (Mb), and the allele frequency of the focal SNP.

and π had a signal at the focal SNP. The Md-rank-*P* compounded these signals into a single measure and captured signals from multiple statistics (the MCD was incorporated into this calculation).

Incorporating the MCD had varying results on the signals of selection, but the results for the empirical data were consistent with the results from the simulations. The MCD generally decreased the performance of M_d, H_d, and N_d, had no effect on DCMS, and resulted in more significant signals for Md-rank-*P* (Fig. S6). For the remaining results, we use the default M_d, the default DCMS, and the MCD with Md-rank-*P*.

For all the compound measures tested, we found that Md-rank-*P* had the best signals of selection with more significant signals (Fig. 6), steeper signals (Figs 7 and S7), and closer signals to the causal SNP (Fig. S8). Although DCMS performed well, Md-rank-*P* outperformed DCMS at measures of steepness (Fig. S9). Generally Md-rank-*P* and DCMS performed well when there was a signal in at least two of the univariate statistics (Figs 6, S7 and S8). On the other hand, CSS performed poorly across all metrics because it did not account for directionality of the *Z*-score. Multivariate distances generally performed poorly on the dog data, regardless of whether the MCD was implemented.

Discussion

Our study took novel steps to compare compound measures for genome scans and explore the utility of robust statistics in the calculation of compound measures. Compound measures can provide an objective criterion to prioritize candidates (in contrast to taking the overlap among different methods) and provide increased resolution to identify selected genomic regions by integrating the signal provided by univariate test statistics (Grossman *et al.* 2010; Ma *et al.* 2015). We found that compound measures could improve the signal of selection and decrease variation in results, though they are not a panacea.

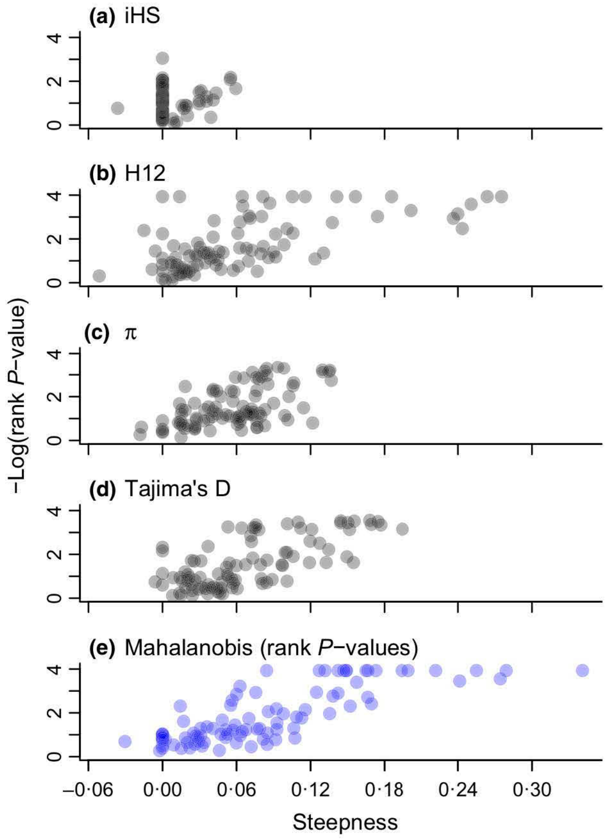

Fig. 7. Significance ($-\log_{10}$ fractional rank) plotted against steepness for each individual locus. Loci in the upper right hand of the graph have a steeper signal of selection and a more significant signal. The univariate statistics are plotted in (a–d), compared to the composite measure Md-rank-*P* plotted in (e).

BEST PRACTICES WHEN USING COMPOSITE MEASURES FOR GENOME SCANS

A single genome scan method may output many variables and the first decision is to choose a variable to represent that method in a composite measure. For example, a genome-wide association study may output an effect size of an SNP, a test statistic for significance of that SNP, and a *P*-value of that test statistic. Ideally, the best variable to choose will measure evidence against an effect size of zero, which is typically the test statistic or the *P*-value. Once the investigator has decided which values to combine, the next decision is which composite measure to apply.

If the goal is to identify outliers regardless of directionality, then a multivariate distance is a good choice for a composite measure. When applying a multivariate distance, Mahalanobis distance or harmonic mean distance should be used if data are parametric and nearest-neighbour distance should be used if data are non-parameteric.

However, for many variables directionality is meaningful (e.g. larger test statistics and smaller *P*-values reflect higher significance), and in this case combining *P*-values would capture this directionality in the composite measure. Note, however, that our demography simulations showed that methods for combining *P*-values may perform worse than multivariate distances when there is high variance in the signal among statistics, as there was in the isolation-by-distance model. When combining *P*-values, the next issue is whether to use *P*-values calculated by the univariate method or to use a rank-based *P*-value. If your data satisfy the assumptions of the statistical method implemented in the program, then the *P*-values will be well calibrated and this is the best case scenario (Fig. S2b). On the other hand, if the univariate method does not calculate *P*-values or makes liberal or conservative assumptions (Fig. S2a,c), then using a rank-based *P*-value would be a better approach (Fig. S2d). These assumptions can be evaluated by both using QQ-plots to check the distribution of *P*-values and by calculating the genomic inflation factor (François *et al.* 2015).

When combining rank-based *P*-values, our analysis of the dog data showed that Md-rank-*P* (with MCD) performed the best of all the statistics, and DCMS performed the next best. The higher performance of Md-rank-*P* over DCMS may have occurred in this case because it is more sensitive to detecting a signal that is significant by only a single univariate statistic, whereas DCMS will tend to reflect signals that are significant by more than one univariate statistic and exhibits odd behaviour depending on the correlation structure of the data (see toy example in Fig. S10). Composite selection signal generally performed poorly because it assumed independence and did not take into account directionality. Take, for example, the combination of a statistic in which a negative signal is indicative of selection (such as Tajima's *D*) with a statistic in which a positive signal is indicative of selection (such as F_{ST}). The former statistic would have a negative *Z*-score and the latter statistic would have a positive *Z*-score, yielding an average *Z*-score for the CSS method near 0, which would not be significant.

Most importantly, all compound measures will reflect the sensitivities of the statistics that they summarize. For instance, extraordinarily large or small values of Tajima's *D* can be caused by changes in population size, and not by selection (Tajima 1989b). While we used Tajima's *D* in this study for the purposes of illustration, investigators should take care when choosing the statistics used in the calculation of the compound measure.

INCORPORATING A ROBUST APPROACH

Despite their apparent utility for identifying genomic outliers, robust statistics have rarely been applied to genome scans. The goal of applying a robust approach is to better capture the neutral expectation, and we found that although the MCD could identify neutral loci in our simulated datasets it did not universally improve the signal of selection in composite measures. Incorporating the MCD typically improved performance of the nearest neighbour distance and Md-rank-*P*, and had no effect on DCMS. In the other cases, incorporating the MCD could improve or decrease performance by as much as 30%. This may have occurred because for many statistics the variation around neutrality is dominated by statistical noise rather than signal, and hence inferring covariances from a too strict a 'robust' set may result in too weak correlations. This may have a profound influence on results and lead to many false positives. Consistent with this reasoning, we found that the MCD generally improved power for Mahalanobis distance in the 1R and 2R models (stronger covariance patterns among loci) and had variable effects on power for the IM and IBD models (weaker covariance patterns among loci). Also consistent with this reasoning, we found that the MCD decreased performance of the multivariate distances in the dog data, where each dataset was based on a single population (weaker covariance patterns among loci). On the other hand, we found that the MCD improved or did not affect power for methods that combined rank *P*-values (Md-rank-*P* and DCMS) in both simulated and empirical data, which suggests that incorporating a robust approach into these methods would be an improvement. While the work presented here has been an important first step into incorporating robust algorithms into the statistical genomics framework, we recommend that investigators evaluate robust approaches for specific scenarios before incorporating them.

Conclusions

Compound measures may provide a tractable, powerful approach for prioritizing candidate regions or loci for further investigation. On the other hand, the univariate statistics that they summarize typically test a specific null hypothesis based on theoretical population genetic models, and therefore offer information about an underlying evolutionary process. Compound measures are naïve to this information and simply infer outliers based on their position in multivariate space. Therefore, it is important for investigators to interpret multivariate outliers in the light of the theoretical models that the univariate statistics are based on, and use functional validation to gather important information on the mode or strength of selection. Given the stochasticity of natural selection and demography, plus the complexities stemming from experimental design, data collection, and univariate analyses, assessing the significance of a compound measure alone is almost certain to remain a difficult task.

Software

The Mahalanobis distance, harmonic mean distance, nearest neighbour distance, CSS, and DCMS measures are implemented in the R package MINOTAUR v1.1.0, which is available via GitHub (https://github.com/NESCent/MINOTAUR) and described in Verity *et al.* 2017. When relevant, the functions also allow users to supply their own covariance matrix or a list of robust points, thereby allowing them to take advantage of robust algorithms.

Authors' contributions

This study was conceptualized by K.E.L. and R.V. K.E.L. performed all analyses and wrote the manuscript. D.C.C. found the empirical dataset and made major contributions to the manuscript. R.V. and C.C. led development of the R package that implemented the compound measures. K.E.L., D.C.C., S.M.S., L.W., C.C., and R.V. all contributed to the development R package and to writing the manuscript.

Acknowledgements

This work began at the Population Genetics in R Hackathon, which was held in March 2015 at the National Evolutionary Synthesis Center (NESCent) in Durham, NC, with the goal of addressing interoperability, scalability and workflow-building challenges for the population genetics package ecosystem in R (R Development Core Team 2016). The authors were participants in the hackathon and are indebted to NESCent (NSF #EF-0905606) for hosting and supporting the event. D.C.C., S.M.S., C.C., and R.V. benefited from travel support to attend this Hackathon. This research has also been supported by funds from Northeastern University to K.E.L. and from the Medical Research Council to R.V.

Data accessibility

The simulated landscape dataset from Lotterhos & Whitlock (2014, 2015) is archived on Dryad (https://doi.org/10.5061/dryad.mh67v). The empirical dog breed dataset from Schlamp *et al.* (2016) is also archived on Dryad (https://doi.org/10.5061/dryad.hf46s). The exact datasets and code used in this study are archived on Dryad (https://doi.org/10.5061/dryad.bp11m; Lotterhos *et al.* 2017).

References

Barrett, R.D.H., Rogers, S.M. & Schluter, D. (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Benjamini, Y. & Hotchberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Brachi, B., Morris, G.P. & Borevitz, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, **12**, 232.

De Mita, S., Thuillet, A.C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. & Vigouroux, Y. (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

De Villemereuil, P.D., Frichot, É., Bazin, É., François, O. & Gaggiotti, O.E. (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Evangelou, E. & Ioannidis, J.P.A. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, **14**, 379–389.

Ewing, G.B. & Jensen, J.D. (2016) The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, **25**, 135–141.

Fariello, M.I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, **193**, 929–941.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, **31**, 1275–1291.

François, O., Martins, H., Caye, K. & Schoville, S.D. (2015) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

Frichot, E. & François, O. (2015) LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot, E., Schoville, S.D., 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.

Garud, N.R., Messer, P.W., Buzbas, E.O. & Petrov, D.A. (2015) Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, **11**, e1005004.

Gautier, M. (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

Grossman, S.R., Shylakhter, I., Karlsson, E.K. *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.

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

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences USA*, **106**, 9362–9367.

Hoban, S., Kelley, J.L., Lotterhos, K.E. *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *American Naturalist*, **188**, 379–397.

Hoekstra, H.E., Hirschmann, R.J., Bundey, R.A., Insel, P.A. & Crossland, J.P. (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Kim, Y. & Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**, 765–777.

Lotterhos, K.E., Card, D.C., Schaal, S.M., Wang, L., Collins, C. & Verity, B. (2017) Data from: Composite measures of selection can improve the signal-—to-—noise ratio in genome scans. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.bp11m.

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

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

Luu, K., Bazin, E. & Blum, M.G.B. (2016) *pcadapt*: an R package to perform genome scans for selection based on principal components analysis. *Molecular Ecology Resources*, **17**, 67–77.

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q. & Simianer, H. (2015) Properties of different selection signature statistics and a new strategy for combining them. *Heredity*, **115**, 426–436.

Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, **2**, 49–55.

Messer, P.W. (2015) H-scan: detecting hard and soft sweeps in population genomic data. Available at: https://messerlab.org/resources/ (accessed 15 January 2017).

Murtagh, F. & Legendre, P. (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, **31**, 274–295.

Nei, M. & Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences USA*, **76**, 5269–5273.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. & Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Pavlidis, P., Zivkovic, D., Stamatakis, A. & Alachiotis, N. (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.

R Development Core Team. (2016) *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org (accessed 15 January 2017).

Randhawa, I.A.S., Khatkar, M.S., Thomson, P.C. & Raadsma, H.W. (2014) Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics*, **15**, 1.

Randhawa, I.A., Khatkar, M.S., Thomson, P.C. & Raadsma, H.W. (2015) Composite selection signals for complex traits exemplified through bovine stature using multibreed cohorts of European and African *Bos taurus*. *G3: Genes| Genomes| Genetics*, **5**, 1391–1401.

Rousseeuw, P.J. & Driessen, K.V. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Savolainen, O., Lascoux, M. & Merila, J. (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schlamp, F., van der Made, J., Stambler, R., Chesebrough, L., Boyko, A.R. & Messer, P.W. (2016) Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, **25**, 342–356.

Schmitt, E., Oellerer, V. & Vakili, K. (2014) The finite sample breakdown point of PCS. *Statistics & Probability Letters*, **94**, 214–220.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & & Williams, R.M.J. (1949) *The American Soldier, Vol. 1: Adjustment During Army Life*. Princeton University Press, Princeton, NJ, USA.

Tajima, F. (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima, F. (1989b) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.

Todorov, V. & Filzmoser, P. (2009) An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, **32**, 1–47.

Togo, L. (2010) *Data Mining With R, Learning With Case Studies*. Chapman and Hall/CRC, New York, NY, USA.

Utsunomiya, Y.T., O'Brien, A.M., Sonstegard, T.S., Van Tassell, C.P., do Carmo, A.S., Meszaros, G., Sölkner, J. & Garcia, J.F. (2013) Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS ONE*, **8**, e64280.

Vakili, K. & Schmitt, E. (2014) Finding multivariate outliers with FastPCS. *Computational Statistics & Data Analysis.*, **69**, 54–66.

Vatsiou, A.I., Bazin, E. & Gaggiotti, O.E. (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology*, **25**, 89–103.

Verity, R., Collins, C., Card, D.C., Schaal, S.M., Wang, L. & Lotterhos, K.E. (2017) MINOTAUR: a platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources*, **17**, 33–43.

Verity, R. & Nichols, R.A. (2014) What is genetic differentiation and how should we measure it - G_{ST}, D, neither or both? *Molecular Ecology*, **23**, 4216–4225.

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biology*, **4**, e72.

Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Whitlock, M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, **18**, 1368–1373.

Whitlock, M.C. & Lotterhos, K.E. (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *American Naturalist*, **186**, S24–S36.

Yang, J., Benyamin, B., McEvoy, B.P. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

Received 8 October 2016; accepted 13 March 2017
Handling Editor: Joanna Kelley

Supporting Information

Details of electronic Supporting Information are provided below.

Data S1. Lotterhos *et al.* 2017 SupplementFigs.pdf contains all supplementary figures.