A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation

Simon Aeschbacher,[*,†,1] Mark A. Beaumont,[‡] and Andreas Futschik[§]

[*]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, [†]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria, [‡]Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TW, United Kingdom, and [§]Institute of Statistics and Decision Support Systems, University of Vienna, 1010 Vienna, Austria

ABSTRACT The choice of summary statistics is a crucial step in approximate Bayesian computation (ABC). Since statistics are often not sufficient, this choice involves a trade-off between loss of information and reduction of dimensionality. The latter may increase the efficiency of ABC. Here, we propose an approach for choosing summary statistics based on boosting, a technique from the machine-learning literature. We consider different types of boosting and compare them to partial least-squares regression as an alternative. To mitigate the lack of sufficiency, we also propose an approach for choosing summary statistics locally, in the putative neighborhood of the true parameter value. We study a demographic model motivated by the reintroduction of Alpine ibex (*Capra ibex*) into the Swiss Alps. The parameters of interest are the mean and standard deviation across microsatellites of the scaled ancestral mutation rate ($\theta_{\mathrm{anc}} = 4N_{\mathrm{e}}u$) and the proportion of males obtaining access to matings per breeding season (ω). By simulation, we assess the properties of the posterior distribution obtained with the various methods. According to our criteria, ABC with summary statistics chosen locally via boosting with the L_2-loss performs best. Applying that method to the ibex data, we estimate $\hat{\theta}_{\mathrm{anc}} \approx 1.288$ and find that most of the variation across loci of the ancestral mutation rate u is between 7.7×10^{-4} and 3.5×10^{-3} per locus per generation. The proportion of males with access to matings is estimated as $\hat{\omega} \approx 0.21$, which is in good agreement with recent independent estimates.

UNDERSTANDING the mechanisms leading to observed patterns of genetic diversity has been a central objective since the beginnings of population genetics (Fisher 1922; Haldane 1932; Wright 1951; Charlesworth and Charlesworth 2010). Three recent trends keep advancing this undertaking: (1) molecular data are becoming available at an ever higher pace (Rosenberg *et al.* 2002; Frazer *et al.* 2007), (2) new theory continues to be developed, and (3) increased computational power allows solution of problems that were intractable just a few years ago. In parallel, the focus has shifted to inference under complex models (*e.g.*, Fagundes *et al.* 2007; Blum and Jakobsson 2011) and to the joint estimation of parameters (*e.g.*, Williamson *et al.* 2005). Usually, these models are stochastic. The increasing complexity of models is justified by the underlying processes like inheritance, mutation, modes of reproduction, and spatial subdivision. On the other hand, complex models are often not amenable to inference based on exact analytical results. Instead, approximate methods such as Markov chain Monte Carlo (MCMC) (Gelman *et al.* 2004) or approximate Bayesian computation (ABC) (Marjoram and Tavaré 2006) are used. These approximate methods address different issues in inference and the choice therefore depends on the specific problem. A significant part of research in the field is currently devoted to the refinement and development of such methods. ABC is a Monte Carlo method of inference that emerged from the confrontation with models for which the evaluation of the likelihood is computationally prohibitive or impossible (Fu and Li 1997; Tavaré *et al.* 1997; Weiss and Von Haeseler 1998; Pritchard *et al.* 1999; Beaumont *et al.* 2002). ABC may be viewed as a class of rejection algorithms (Marjoram *et al.* 2003; Marjoram and Tavaré 2006), where the full data are projected to a lower-dimensional set of summary statistics. Here, we propose an approach for choosing summary statistics based on boosting (see below), and we apply it to the estimation of the mean and variance

doi: 10.1534/genetics.112.143164
Manuscript received June 23, 2012; accepted for publication August 22, 2012
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.143164/-/DC1.

[1]Corresponding author: Faculty of Mathematics, University of Vienna, Nordbergstr. 15, A-1090 Vienna, Austria. E-mail: simon.aeschbacher@univie.ac.at

across microsatellites of the scaled ancestral mutation rate and of the mating skew in Alpine ibex (*Capra ibex*). We further show that focusing the choice of statistics on the putative neighborhood of the true parameter value improves estimation in this context.

The principle of ABC is to first simulate data under the model of interest and then accept simulations that produced data close to the observation. Parameter values belonging to accepted simulations yield an approximation to the posterior distribution, without the need to explicitly calculate the likelihood. The full data are usually compressed to summary statistics to reduce the number of dimensions. Formally, the posterior distribution of interest is given by

$$\pi(\boldsymbol{\phi}|D) = \frac{\pi(D|\boldsymbol{\phi})\pi(\boldsymbol{\phi})}{\pi(D)} = \frac{\pi(D|\boldsymbol{\phi})\pi(\boldsymbol{\phi})}{\int_{\Phi}\pi(D|\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi}}, \tag{1}$$

where $\boldsymbol{\phi}$ is a vector of parameters living in space Φ, D denotes the observed data, $\pi(\boldsymbol{\phi})$ the prior distribution, and $\pi(D\,|\,\boldsymbol{\phi})$ the likelihood. With ABC, (1) is approximated by

$$\pi_{\varepsilon}(\boldsymbol{\phi}|\mathbf{s}) \propto \pi(\rho(\mathbf{s}', \mathbf{s}) \leq \delta_{\varepsilon}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}), \tag{2}$$

where $\mathbf{s}$ and $\mathbf{s}'$ are abbreviations for realizations of $\mathbf{S}(D)$ and $\mathbf{S}(D')$, respectively, and $\mathbf{S}$ is a function generating a q-dimensional vector of summary statistics calculated from the full data. The prime denotes simulated points, in contrast to quantities related to the observed data. Further, $\rho(\cdot)$ is a distance metric and δ_{ε} the rejection tolerance in that metric space, such that on average a proportion ε of all simulated points is accepted. ABC, its position in the ensemble of model-based inference methods, and its application in evolutionary genetics are reviewed in Marjoram *et al.* (2003), Beaumont and Rannala (2004), Marjoram and Tavaré (2006), Beaumont (2010), Bertorelle *et al.* (2010), and Csilléry *et al.* (2010). Although the origin of ABC is generally assigned to Fu and Li (1997), Tavaré *et al.* (1997), and Pritchard *et al.* (1999), some aspects, such as the summary description of the full data, inference for implicit stochastic models, and algorithms directly sampling from the posterior distribution, trace farther back (*e.g.*, Diggle 1979; Diggle and Gratton 1984; Rubin 1984).

A fundamental issue with the basic ABC rejection algorithm (*e.g.*, Marjoram *et al.* 2003) is its inefficiency: A large number of simulations are needed to obtain a satisfactory number of accepted runs. This problem becomes worse as the number of summary statistics increases and is known as the curse of dimensionality. Three solutions have been proposed: (1) more efficient algorithms combining ABC with principles of MCMC (*e.g.*, Marjoram *et al.* 2003; Wegmann *et al.* 2009) or sequential Monte Carlo (*e.g.*, Sisson *et al.* 2007, 2009; Beaumont *et al.* 2009; Toni *et al.* 2009); (2) fitting a statistical model to describe the relationship of parameters and summary statistics after the rejection step, allowing for a larger tolerance δ_{ε} (Beaumont *et al.* 2002; Blum and François 2010; Leuenberger and Wegmann 2010); and (3) reduction of dimensions by sophisticated choice of summary statistics (*e.g.*, Joyce and Marjoram 2008; Wegmann *et al.* 2009). In this study, we focus on point 3, which involves two further issues. First, most summary statistics used in evolutionary genetics are not sufficient. A summary statistic $S(D)$ is sufficient for parameter ϕ if the conditional probability distribution of the full data D, given $S(D)$ and ϕ, does not depend on ϕ, *i.e.*, if

$$\pi(D = d|S(D) = s, \phi) = \pi(D = d|S(D) = s). \tag{3}$$

In other words, a statistic is sufficient for a parameter of interest, if it contains all the information on that parameter that can possibly be extracted from the full data (*e.g.*, Shao 2003). Second, the choice of summary statistics implies the choice of a suitable metric $\rho(\cdot)$ to measure the "closeness" of simulations to observation (except for the nongeneric case $\varepsilon = 0$ in which no metric needs to be defined). The Euclidean distance (or a weighted version, *e.g.*, Hamilton *et al.* 2005) has been used in most applications, but it is not obvious why this should be optimal. By "optimal" we mean that the resulting posterior estimate performs best in terms of an error criterion (or a set of criteria). The Euclidean distance is a scale-dependent measure of distance—changing the scale of measurement changes the results. Since this scale is determined by the summary statistics, the choice of summary statistics has implications for the choice of the metric. For these reasons, the choice of summary statistics should aim at reducing the dimensions, but also at extracting (combinations of) statistics that contain the essential information about the parameters of interest. This task is reminiscent of the classical problem of variable selection in statistics and machine learning (Hastie *et al.* 2011), and it is of principal interest here.

The choice of summary statistics in ABC has become a focus of research only recently. Joyce and Marjoram (2008) proposed a sequential scheme based on the principle of approximate sufficiency. Statistics are included if their effect on the posterior distribution is larger than some threshold. Their approach seems demanding to implement, and it is not obvious how to define an optimal threshold. Wegmann *et al.* (2009) suggested partial least-squares (PLS) regression as an alternative. In this context, PLS regression seeks linear combinations of the original summary statistics that are maximally decorrelated and, at the same time, have high correlation with the parameters (Hastie *et al.* 2011). A reduction in dimensions is achieved by choosing only the first r PLS components, where r is determined via cross-validation. PLS is one of several approaches for variable selection, but it is an open question how it compares to alternative methods in any specific ABC setting. Moreover, the optimal choice of summary statistics may depend on the location of the true (but unknown) parameter values. By definition, this is to be expected whenever the summary statistics are not sufficient, because then the information extracted from the full data by the summary statistics depends on the parameter value (see Equation 3). It is therefore not obvious why methods that assess the relation between statistics and parameters on a global scale should be optimal. Instead, focusing on the correlation only in the (supposed) neighborhood of the true parameter values might be preferable. The issue is that this neighborhood is not known in advance—if we could choose an arbitrarily small neighborhood around the truth, our inference problem would be solved and we would not need ABC or any other approximate method. However, the neighborhood may be established approximately, as we will argue later. The idea of focusing the choice of summary statistics on some local optimization has also been followed by Nunes and Balding (2010) and Fearnhead and Prangle (2012). Nunes and Balding (2010) proposed using a minimum-entropy algorithm to identify the neighborhood of the true value and then chose the set of summary statistics that minimized the mean squared error across a test data set. Fearnhead and Prangle (2012), on the other hand, first proved that, for a given loss function, an *optimal* summary statistic may be defined. For example, when the quadratic loss is used to quantify the cost of an error, the optimal summary statistic is the posterior mean. Since the latter is not available *a priori*, the authors devised a heuristic to estimate it and were able to show good performance of their approach. The choice of the optimization criterion may include a more local or a global focus on the parameter range. Different criteria will lead to different optimal summary statistics. The approaches by Nunes and Balding (2010) and Fearnhead and Prangle (2012), and the one we take here, have in common that they employ a two-step procedure, first defining "locality" and then using standard methods from statistics or machine learning to select summary statistics in this restricted range. They differ in the details of these two steps (see *Discussion*).

Here, we propose a novel approach for choosing summary statistics in ABC. It is based on boosting, a method developed in machine learning to establish the relationship between predictors and response variables in complex models (Schapire 1990; Freund 1995; Freund and Schapire 1996, 1999). Given some training data, the idea of boosting is to iteratively train a function that describes this relationship. At each iteration, the training data are reweighted according to the current prediction error (loss), and the function is updated according to an optimization rule. It has been argued that boosting is relatively robust to overfitting (Friedman *et al.* 2000), which would be an advantage with regard to high-dimensional problems as encountered in ABC. Different flavors of boosting exist, depending on assumptions about the error distribution, the loss function, and the learning procedure. In a simulation study, we compare the performance of ABC with three types of boosting to ABC with summary statistics chosen via PLS and to ABC with all candidate statistics. We further suggest an approach for choosing summary statistics locally and compare the local variants of the various methods to their global versions. Throughout, we study a model that is motivated by the reintroduction of Alpine ibex into the Swiss Alps. The parameters of interest are the mean and standard deviation across microsatellites of the scaled ancestral mutation rate and the proportion of males that obtain access to matings per breeding season. This model is used first in the simulation study for inference on synthetic data and assessment of performance. Later, we apply the best method to infer posterior distributions given genetic data from Alpine ibex. It is not our goal to compare all the approaches recently proposed for choosing summary statistics in ABC. This would reach beyond the scope of this article, but provides a perspective for future research. Recently, Blum *et al.* (2012) carried out a comparative study of the various approaches and found that, for an example similar to our context, PLS performed slightly better than approximate sufficiency (Joyce and Marjoram 2008), but worse than a number of alternative approaches including the posterior loss method (Fearnhead and Prangle 2012) and the two-stage minimum entropy procedure (Nunes and Balding 2010). Nevertheless, PLS has been widely used in recent applications and we have therefore focused on comparing our approach to PLS.

We start by describing the ibex model and its parameters. We then present an ABC algorithm that includes a step for choosing summary statistics. Later, we describe the boosting approach for choosing the statistics and we suggest how to focus this choice on the putative neighborhood of the true parameter value. Comparing different versions of boosting among each other and with PLS, we conclude that boosting with the L_2-loss restricted to the vicinity of the true parameter performs best, given our criteria. However, the difference from the next best methods (local boosting with the L_1-loss and local PLS) is small.

Model and Parameters

We study a neutral model of a spatially structured population with genetic drift, mutation, and migration. The demography includes admixture, subdivision, and changes in population size. This model is motivated by the recent history of Alpine ibex and their reintroduction into the Swiss Alps (Figures 1 and 2). By the beginning of the 18th century, Alpine ibex had been extinct except for $\sim$100 individuals in the Gran Paradiso area in Northern Italy (Figure 1). At the beginning of the 20th century, a schedule was set up to reestablish former demes in Switzerland (Couturier 1962; Stuwe and Nievergelt 1991; Scribner and Stuwe 1994; Maudet *et al.* 2002). The reintroduction has been documented in great detail by game keepers and authorities. We could reconstruct for 35 demes their census sizes between 1906 and 2006 (Supporting Information, File S2, census sizes) and the number of females and males transferred between them, as well as the times of these founder/

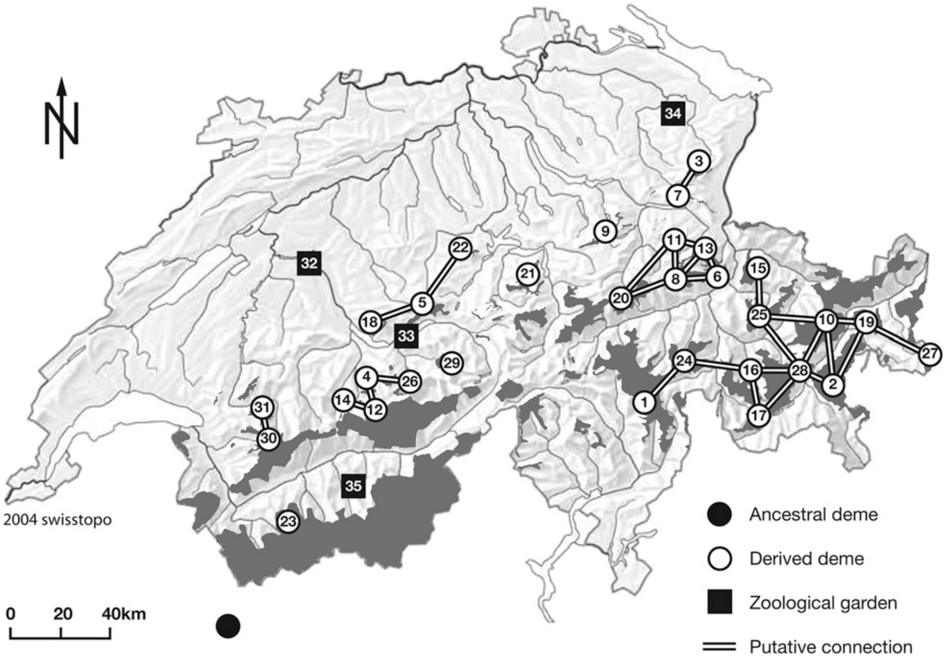


Figure 1 Location of Alpine ibex demes in the Swiss Alps. The parts with dark shading represent areas inhabited by ibex. The ancestral deme is located in the Gran Paradiso area in Northern Italy, close to the Swiss border. The two demes in the zoological gardens 33 and 34 were first established from the ancestral one. Further demes, including the two in zoological gardens 32 and 35, were derived from demes 33 and 34. Putative connections indicate the pairs of demes for which migration is considered possible. For a detailed record of the demography and the genealogy of demes see Figure S1 and File S3. For deme names see Table S1. The map was obtained via the Swiss Federal Office for the Environment (FOEN) and modified with permission.

admixture events (File S3, transfers). Inference on mutation and migration can therefore be done conditional on this information. The signal for this inference comes from the distribution of allele frequencies across loci and across demes.

We constructed a forward-in-time model starting with an ancestral gene pool d_{anc} of unknown effective size, N_{e}, representing the Gran Paradiso ibex deme. At times t_1 and t_2, two demes, d_1 and d_2, are derived from the ancestral gene pool. They represent the breeding stocks that were established in two zoological gardens in Switzerland in 1906 and 1911 (Figure 1) (Stuwe and Nievergelt 1991). Further demes are then derived from these. We let t_i be the time at which deme d_i is established. Once a derived deme has been established, it may contribute to the foundation of additional demes. The sizes of derived demes follow the observed census size trajectories (File S2, census sizes). We interpolated missing values linearly, if the gap was only 1 year, or exponentially, if values for ≥2 successive years were missing. Derived demes may exchange migrants if they are connected. This depends on information obtained from game keepers and on geography (Figure 1). Given a pair of connected demes d_i and d_j, we define the forward migration rates, $\tilde{m}_{i,j}$ and $\tilde{m}_{j,i}$. More precisely, $\tilde{m}_{i,j}$ is the proportion of potential emigrants (see File S1) in deme d_i that migrate to deme d_j per year. We assume that $\tilde{m}_{i,j}$ is constant over time and the same for females and males. Migration is included in the model, although we do not estimate migration rates in this article, but in a related article (S. Aeschbacher, A. Futschik, and M. A. Beaumont, unpublished results). Here, we restrict our attention to the ancestral mutation rate and the proportion of males getting access to matings, marginal to the migration rates (see below). Estimating migration rates comes with additional complications that go beyond the focus of this article. A schematic representation of the model is given in Figure 2. When modeling migration, reproduction, and founder events, we take into account the age structure of the population (see File S1 for details).

Population history is split into two phases. The first started at some unknown point in the past and ended at $t_i = 1906$, when the first ibex were brought from Gran Paradiso (d_{anc}) to d_1. For this ancestral phase, we assume constant, but unknown effective size N_{e} and mutation following the single stepwise model (Ohta and Kimura 1973) at a rate u per locus and generation. Accordingly, we define the scaled mutation rate in the ancestral deme as $\theta_{\text{anc}} = 4N_{\text{e}}u$. Mutation rates may vary among microsatellites for several reasons (Estoup and Cornuet 1999). To account for this, we use a hierarchical model, assuming that θ_{anc} is normally distributed across loci on the $\log_{10}$-scale with mean $\mu_{\theta_{\text{anc}}}$ and standard deviation $\sigma_{\theta_{\text{anc}}}$. In our case, $\mu_{\theta_{\text{anc}}}$ and $\sigma_{\theta_{\text{anc}}}$ are the hyperparameters (Gelman *et al.* 2004) of interest. We assume that N_{e} is the same for all loci, so that variance in θ_{anc} can be attributed to u exclusively. In principle, variation in diversity across loci could also be due to selection at linked genes (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993; Barton 2000), rather than variable mutation rates. Most likely, we cannot distinguish these alternatives with our data. The second, recent phase started at time t_1 and went up to the time of genetic sampling, $t_{\text{g}} = 2006$. During this phase, the numbers of males and females transferred at founder/admixture events and census population sizes are known and accounted for. Mutation is neglected, since, in the case of ibex, this phase spans only ~11 generations at most (Stuwe and Grodinsky 1987). At the transition from the ancestral to the recent phase, genotypes of the founder individuals introduced to demes d_1 and d_2 are sampled at

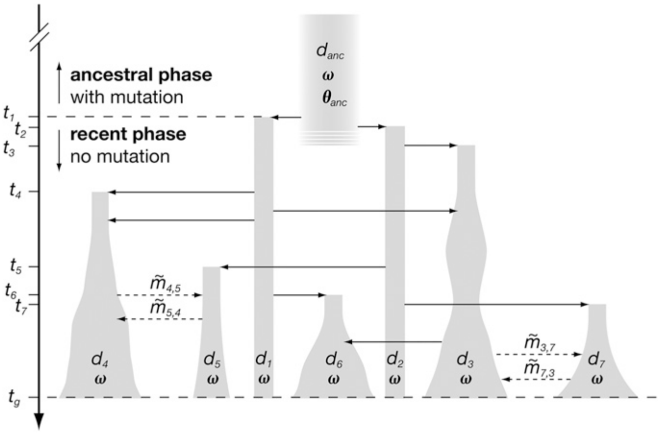


Figure 2 Schematic representation of the demographic model motivated by the reintroduction of Alpine ibex into the Swiss Alps. Shaded shapes represent demes, indexed by d_i, and the width of the shapes reflects the census size. Time goes forward from top to bottom, and the point in time when deme d_i is established is shown as t_i; t_g is the time of genetic sampling. The total time is split by t_1 into an ancestral phase with mutation and a recent phase for which mutation is ignored (see text for details). Solid horizontal arrows represent founder/admixture events and dashed arrows migration. The parameters are (1) the scaled mutation rate in the ancestral deme, $\theta_{\text{anc}} = 4N_eu$; (2) the proportion of males getting access to matings, ω; and (3) forward migration rates between putatively connected demes, $\tilde{m}_{i,j}$. The actual model considered in the study contains 35 derived demes (Figure 1 and Table S1). The exact demography is reported in Figure S1 and File S3, transfers.

random from the ancestral deme, d_{anc}. At the end of the recent phase (t_g), genetic samples are taken according to the sampling scheme under which the real data were obtained. Of the total 35 demes, 31 were sampled (Table S1).

In Alpine ibex, male reproductive success is highly skewed toward dominant males. Dominance is correlated with male age (Willisch *et al.* 2012), and ranks are established during summer. Only a small proportion of males obtain access to matings during the rut in winter (Aeschbacher 1978; Stuwe and Grodinsky 1987; Scribner and Stuwe 1994; Willisch and Neuhaus 2009; Willisch *et al.* 2012). To take this into account, we introduce the proportion of males obtaining access to matings, ω, as a parameter. It is defined relative to the number of potentially reproducing males (and therefore conditional on male age; see File S1) and has an impact on the strength of genetic drift. For simplicity, we assume that ω is the same in all demes and independent of deme size and time.

In principle, we want to infer the joint posterior distribution $\pi(\tilde{\mathbf{m}}, \boldsymbol{\alpha}\,|\,D)$, where $\boldsymbol{\alpha} = (\mu_{\theta_{\text{anc}}}, \sigma_{\theta_{\text{anc}}}, \omega)$ and $\tilde{\mathbf{m}} = \{\tilde{m}_{i,j} : i \neq j, i \in \mathcal{J}_m,\ j \in \mathcal{J}_m\}$, with $\mathcal{J}_m$ denoting the set of all demes connected via migration to at least one other deme (Figure 1). This is a complex problem because there are many parameters and even more candidate summary statistics; the curse of dimensionality is severe. Targeting the joint posterior with ABC naively would give a result, but it would be hard to assess its validity. It is more promising to address intermediate steps and assess them one by one. A first step is to focus on a subset of parameters and marginalize over the others. By marginalizing we mean that the joint posterior distribution is integrated with respect to the parameters that are not of interest. In our case, we may focus on $\boldsymbol{\alpha}$ and integrate over the migration rates $\tilde{\mathbf{m}}$ where they have prior support (Table 1). In practice, marginal posteriors can be targeted directly with ABC—without the need to compute the joint likelihood explicitly and integrate over it (see below). A second step is to clarify what summary statistics should be chosen for the subset of focal parameters ($\boldsymbol{\alpha}$). A third one is to deal with the curse of dimensionality related to estimating $\tilde{\mathbf{m}}$. In this article, we deal with steps one and two: We aim at estimating $\boldsymbol{\alpha}$ marginally to $\tilde{\mathbf{m}}$ and we seek a good method for choosing summary statistics with respect to $\boldsymbol{\alpha}$. The third step—estimating $\tilde{\mathbf{m}}$ and dealing with its high dimensionality—is treated separately (S. Aeschbacher, A. Futschik, and M. A. Beaumont, unpublished results). Note that this division of the problem implies the assumption that priors of the migration rates and male mating success are independent. We make this assumption partly for convenience and partly because we are not aware of any study that has shown a relation between the two in Alpine ibex. The division into two steps also requires that the set of all summary statistics ($\mathbf{S}$) can be split into two subsets, such that the first ($\mathbf{S}_{\boldsymbol{\alpha}}$) contains most of the information on $\boldsymbol{\alpha}$, whereas the second ($\mathbf{S}_{\tilde{\mathbf{m}}}$) contains most of the information on $\tilde{\mathbf{m}}$. Moreover, $\mathbf{S}_{\boldsymbol{\alpha}}$ should not be affected much by $\tilde{\mathbf{m}}$. As shown in the *Appendix*, the results are not much affected in such a situation while the computational burden decreases significantly. The arguments in the *Appendix* rely on the notions of approximate sufficiency and approximate ancillarity.

Methods

The joint posterior distribution of our model may be factorized as

$$\pi\big(\tilde{\mathbf{m}}, \boldsymbol{\alpha}|D\big) = \pi\big(\tilde{\mathbf{m}}|\boldsymbol{\alpha}, D\big)\pi(\boldsymbol{\alpha}|D). \tag{4}$$

As mentioned, here we target only the marginal posterior of $\boldsymbol{\alpha}$, which is formally obtained as

$$\pi(\boldsymbol{\alpha}|D) = \int_{\mathcal{M}} \pi\big(\tilde{\mathbf{m}}, \boldsymbol{\alpha}|D\big)d\tilde{\mathbf{m}}, \tag{5}$$

where $\mathcal{M}$ is the domain of possible values for $\tilde{\mathbf{m}}$. By the nature of our problem, $\pi(\tilde{\mathbf{m}}, \boldsymbol{\alpha}\,|\,D)$ is not available. However, with ABC we may target (5) directly by sampling from $\pi_\varepsilon(\boldsymbol{\alpha}\,|\,\mathbf{s}_{\boldsymbol{\alpha}} = \mathbf{S}_{\boldsymbol{\alpha}}(D))$, where we assume that $\mathbf{S}_{\boldsymbol{\alpha}}$ is a subset of summary statistics approximately sufficient for estimating $\boldsymbol{\alpha}$ (*Appendix*). Note that $\mathbf{S}_{\boldsymbol{\alpha}}$ may not be sufficient to estimate the joint posterior (4), however (Raiffa and Schlaifer 1968). The following standard ABC algorithm provides an approximation to $\pi(\boldsymbol{\alpha}\,|\,\mathbf{s}_{\boldsymbol{\alpha}})$ (*e.g.*, Marjoram *et al.* 2003):

Table 1 Parameters and prior distributions

Parameter	Description	Prior distribution
$\theta_{\text{anc},l}$	Scaled ancestral mutation rate at locus l, $4N_e u$	$\log_{10}(\theta_{\text{anc},l}) \sim N(\mu_{\theta_{\text{anc}}}, \sigma^2_{\theta_{\text{anc}}})^a$
$\mu_{\theta_{\text{anc}}}$	Mean across loci of $\theta_{\text{anc},l}$ (on $\log_{10}$-scale)	$\mu_{\theta_{\text{anc}}} \sim N(0.5, 1)$
$\sigma_{\theta_{\text{anc}}}$	Standard deviation across loci of $\theta_{\text{anc},l}$ (on $\log_{10}$-scale)	$\sigma_{\theta_{\text{anc}}} \sim \log_{10}$-uniform in [0.01, 1]
ω	Proportion of mature males with access to matings	$\omega \sim \log_{10}$-uniform in [0.01, 1]
$\tilde{m}_{i,j}{}^b$	Forward migration rate per year from deme i to deme j	$\tilde{m}_{i,j} \sim \log_{10}$-uniform in $[10^{-3.5}, 10^{-0.5}]$

[a] $N(\mu, \sigma^2)$, normal distribution with mean μ and variance σ^2.
[b] Although migration rates are not estimated here, they are drawn from the prior in all simulations (see main text).

Algorithm A

```
A1. Calculate summary statistics $\mathbf{s_\alpha} = \mathbf{S_\alpha}(D)$ from observed data.
A2. For $t = 1$ to $t = N$,
    i. Sample $(\boldsymbol{\alpha}'_t, \tilde{\mathbf{m}}'_t)$ from $\pi(\boldsymbol{\alpha}, \tilde{\mathbf{m}}) = \pi(\boldsymbol{\alpha})\pi(\tilde{\mathbf{m}})$.
    ii. Simulate data $D'_t$ (at all loci and for all demes) from $\pi(D \mid \boldsymbol{\alpha}'_t,\ \tilde{\mathbf{m}}'_t)$.
    iii. Calculate $\mathbf{s}'_{\boldsymbol{\alpha},t} = \mathbf{S_\alpha}(D'_t)$ from simulated data.
A3. Scale $\mathbf{s_\alpha}$ and $\mathbf{s}'_{\boldsymbol{\alpha},t}$ $(t = 1, \ldots, N)$ appropriately.
A4. For each $t$, accept $\boldsymbol{\alpha}'_t$ if $\rho(\mathbf{s}'_{\boldsymbol{\alpha},t}, \mathbf{s_\alpha}) \le \delta_\varepsilon$, using scaled summary statistics from A3.
A5. Estimate the posterior density $\pi_\varepsilon(\boldsymbol{\alpha} \mid \mathbf{s_\alpha})$ from the $\varepsilon N$ accepted points $\langle \mathbf{s}'_{\boldsymbol{\alpha},t}, \boldsymbol{\alpha}'_t \rangle$.
```

Step A2 may be easily parallelized on a cluster computer. In doing so, one needs to store $\langle \mathbf{s}'_{\boldsymbol{\alpha},t},\ \boldsymbol{\alpha}'_t \rangle$. Step A5 may include postrejection adjustment via regression (Beaumont *et al.* 2002; Blum and François 2010; Leuenberger and Wegmann 2010) and scaling of parameters. In general, the set of well-chosen, informative summary statistics $\mathbf{S_\alpha}$ is not known in advance. Instead, a set of candidate statistics $\mathbf{S}$ (chosen based on intuition or analogy to simpler models) may be available. Therefore, we propose algorithm B—a modified version of algorithm A—that includes an additional step for the empirical choice of summary statistics $\mathbf{S_\alpha}$ informative on $\boldsymbol{\alpha}$ given a set of candidate statistics, $\mathbf{S}$ (for similar approaches, see Hamilton *et al.* 2005; Wegmann *et al.* 2009):

Algorithm B

```
B1. Calculate candidate summary statistics $\mathbf{s} = \mathbf{S}(D)$ from observed data.
B2. For $t = 1$ to $t = N$,
    i. Sample $(\boldsymbol{\alpha}'_t, \tilde{\mathbf{m}}'_t)$ from $\pi(\boldsymbol{\alpha}, \tilde{\mathbf{m}}) = \pi(\boldsymbol{\alpha})\pi(\tilde{\mathbf{m}})$.
    ii. Simulate data $D'_t$ (at all loci and for all demes) from $\pi(D|\boldsymbol{\alpha}'_t, \tilde{\mathbf{m}}'_t)$.
    iii. Calculate candidate summary statistics $\mathbf{s}'_t = \mathbf{S}(D'_t)$ from simulated data.
B3. Sample without replacement $n \le N$ simulated pairs $\langle \mathbf{s}'_t, \boldsymbol{\alpha}'_t \rangle$, denote them by $\langle \mathbf{s}'_{t^*}, \boldsymbol{\alpha}'_{t^*} \rangle$, and use them as a training data set to choose informative statistics $\mathbf{S_\alpha}$.
B4. According to B3, obtain $\mathbf{s_\alpha}$ from $\mathbf{s}$; for $t = 1$ to $t = N$, obtain $\mathbf{s}'_{\boldsymbol{\alpha},t}$ from $\mathbf{s}'_t$.
B5. Scale $\mathbf{s_\alpha}$ and $\mathbf{s}'_{\boldsymbol{\alpha},t}$ $(t = 1, \ldots, N)$ appropriately.
B6. For each $t$, accept $\boldsymbol{\alpha}'_t$ if $\rho(\mathbf{s}'_{\boldsymbol{\alpha},t}, \mathbf{s_\alpha}) \le \delta_\varepsilon$, using scaled summary statistics from B5.
B7. Estimate the posterior density $\pi_\varepsilon(\boldsymbol{\alpha} \mid \mathbf{s_\alpha})$ from the $\varepsilon N$ accepted points $\langle \mathbf{S}'_{\boldsymbol{\alpha},t}, \boldsymbol{\alpha}'_t \rangle$.
```

Note that $\mathbf{S_\alpha}$ in steps B3 and B4 may be either a subset of $\mathbf{S}$ or some function (*e.g.*, a linear combination) of $\mathbf{S}$ (details of implementation given below). In the following, we describe a novel approach based on boosting and recently proposed by Lin *et al.* (2011) for the choice of $\mathbf{S_\alpha}$ in B3.

Choice of summary statistics via boosting

Boosting is a collective term for meta-algorithms originally developed for supervised learning in classification problems (Schapire 1990; Freund 1995). Later, versions for regression (Friedman *et al.* 2000) and other contexts were developed (Bühlmann and Hothorn 2007 and references therein). Assume a set of n observations indexed by i and associated with a one-dimensional response Y_i. For (binary) classification, $Y_i \in \{0, 1\}$, but in a regression context, Y_i may be continuous in $\mathbb{R}$. Further, each observation is associated with a vector of q predictors $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(q)})$. Given a training data set $\{\langle \mathbf{X}_1, Y_1 \rangle, \ldots, \langle \mathbf{X}_n, Y_n \rangle\}$, the task of a boosting algorithm is to learn a function $F(\mathbf{X})$ that predicts Y. Boosting was invented to deal with cases where the relationship between predictors and response is potentially complex, for example, nonlinear (Schapire 1990; Freund 1995; Freund and Schapire 1996, 1999). Establishing the relationship between predictors and response, and weighting predictors according to their importance, directly relates to the problem of choosing summary statistics in ABC: Given candidate statistics $\mathbf{S}$, we want to find a subset or combination of statistics $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ informative for the kth parameter $\boldsymbol{\alpha}^{(k)}$ in $\boldsymbol{\alpha}$, for every k. Taking the set of simulated pairs $\langle \mathbf{s}'_t, f(\boldsymbol{\alpha}'_t) \rangle$ $(t = 1, \ldots, N)$ from step B3 of algorithm B as a training data set, this may be achieved by boosting. For this purpose, we interpret the summary statistics $\mathbf{S}$ as predictors $\mathbf{X}$ and the parameters $\boldsymbol{\alpha}^{(k)}$ as the response Y. Note that we use $f(\boldsymbol{\alpha}'_t)$ to be generic in the sense that the response might actually be a function—such as a discretization step (see below)—of $\boldsymbol{\alpha}'_t$.

The principle of boosting is to iteratively apply a *weak learner* to the training data and then combine the ensemble of weak learners to construct a *strong learner.* While the weak learner predicts only slightly better than random guessing, the strong learner will usually be well correlated with the true Y. This is because the training data are reweighted after each step according to the current error, such that the next weak learner will focus on those observations that were particularly hard to assign. However, too strong a correlation will lead to overfitting, so that in practice one defines an upper limit for the number of iterations (see below). The behavior of the weak learner is described by the base procedure $\hat{g}(\cdot)$, a real valued function. The final result (strong learner) is the desired function estimate $\hat{F}(\cdot)$. Given a loss function $L(\cdot,\cdot)$ that quantifies the disagreement between Y and $F(\mathbf{X})$, we want to estimate the function that minimizes the expected loss,

$$F^*(\cdot) = \arg\min_{F(\cdot)}\mathbb{E}[L(Y, F(\mathbf{X}))]. \tag{6}$$

This can be done by considering the empirical risk $n^{-1}\sum_{i=1}^{n}L(Y_i, F(\mathbf{X}_i))$ and pursuing iterative steepest descent in function space (Friedman 2001; Bühlmann and Hothorn 2007). The corresponding algorithm is given in the *Appendix*. The generic boosting estimator obtained from this algorithm is a sum of base procedure estimates,

$$\hat{F}(\cdot) = \nu\sum_{m=1}^{m_{\text{stop}}}\hat{g}^{[m]}(\cdot). \tag{7}$$

Both ν and m_{stop} are tuning parameters that essentially control the overfitting behavior of the algorithm. Bühlmann and Hothorn (2007) argue that the learning rate ν is of minor importance as long as $\nu \leq 0.1$. The number of iterations, m_{stop}, however, should be chosen specifically in any application via cross-validation, bootstrapping, or some information criterion [*e.g.*, Akaike's information criterion (AIC)].

Base procedure: Different versions of boosting are obtained depending on the base procedure $\hat{g}(\cdot)$ and the loss function $L(\cdot,\cdot)$. Here, we let $\hat{g}(\cdot)$ be a simple componentwise linear regression (Bühlmann and Hothorn 2007; see *Appendix*). With this choice, the boosting algorithm selects in every iteration only one predictor, namely the one that is most effective in reducing the current loss. For instance, with the L_2-loss (defined below), after each step, $\hat{F}(\cdot)$ is updated linearly according to

$$\hat{F}^{[m]}(\mathbf{x}) = \hat{F}^{[m-1]}(\mathbf{x}) + \nu\hat{\lambda}^{(\hat{\zeta}_m)}\mathbf{x}^{(\hat{\zeta}_m)}, \tag{8}$$

where $\hat{\zeta}_m$ denotes the index of the predictor variable selected in iteration m. Accordingly, in iteration m only the $\hat{\zeta}$th component of the coefficient estimate $\hat{\lambda}^{[m]}$ is updated. As m goes to infinity, $\hat{F}(\cdot)$ converges to a least-squares solution. In practice, we stop at m_{stop}, and we denote the final vector of estimated coefficients as $\hat{\lambda} = \hat{\lambda}^{[m_{\text{stop}}]}$. Recall that in our context, the predictor variables $\mathbf{X}$ correspond to the candidate summary statistics $\mathbf{S}$. For each of the k parameters in $\boldsymbol{\alpha}$, we estimate one function $\hat{F}^{[m_{\text{stop}}]}$ and use it to obtain new parameter-specific statistics $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$.

Loss functions: We employed boosting with three loss functions. The first two, L_1-loss and L_2-loss, are appropriate for a regression context with a continuous response $Y \in \mathbb{R}$. In this case, the parameters $\boldsymbol{\alpha}_t'$ are directly interpreted as y_i [*i.e.*, $f(\boldsymbol{\alpha}_t') = \boldsymbol{\alpha}_t'$]. The L_1-loss is given by

$$L_{L_1}(y,F) = |y - F| \tag{9}$$

and results in L_1Boosting. The L_2-loss is given by

$$L_{L_2}(y,F) = \frac{1}{2}|y - F^2| \tag{10}$$

and results in L_2Boosting. The scaling factor $\frac{1}{2}$ in (10) ensures that the negative gradient vector U in the functional gradient descent (FGD) algorithm (*Appendix* and File S1) equals the residuals (Bühlmann and Hothorn 2007). L_1- and L_2Boosting result in a fit of a linear regression, similarly to ordinary regression using the least absolute deviation (L_1-norm) or the least-squares criterion (L_2-norm), respectively. The difference, and a potential advantage of boosting, is that residuals are fitted multiple times depending on the importance of the components of $\mathbf{X}$. Moreover, boosting is considered less prone to overfitting than ordinary L_1- or L_2-fitting (Bühlmann and Hothorn 2007). In general, the L_1-loss is more robust to outliers, but it may produce multiple, potentially unstable solutions. Using L_1- and L_2Boosting to choose summary statistics means assuming a linear relationship between summary statistics and parameters. This is a strong assumption and most likely not globally true. However, the advantage is that the resulting linear combination has only one dimension, such that the curse of dimensionality in ABC may be strongly reduced. Again, the approach using the L_1- or L_2-loss results in one linear combination $\hat{F}^{[m_{\text{stop}}]}$ per parameter $\boldsymbol{\alpha}^{(k)}$, such that $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ has only one component. These linear combinations may end up being correlated across parameters, especially if parameters are not identifiable, *e.g.*, because they are confounded with each other.

To motivate the third loss function, we propose considering the choice of summary statistics as a classification problem. Imagine two classes of parameter values—say, high values in one class and low values in the other. We may ask what summary statistics are important to assign simulations to one of these two classes. With $Y \in \{0, 1\}$ as the class label and $p(\mathbf{x}) := \Pr[Y = 1|\mathbf{X} = \mathbf{x}]$, a natural choice is the negative binomial log-likelihood loss

$$L_{\text{log-lik}}\,(y,p) = -[y\log(p) + (1-y)\log(1-p)], \tag{11}$$

omitting the argument of p for ease of notation. If we parameterize $p = e^F/(1 + e^F)$ so that we obtain $F = \log[p/(1-p)]$ corresponding to the logit-transformation, the loss in (11) becomes

$$L_{\text{log-lik}}(y,F) = \log\left[1 + e^{-(2y-1)F}\right]. \tag{12}$$

The corresponding boosting algorithm is called LogitBoost (or binomial boosting) (Bühlmann and Hothorn 2007). An advantage is that it does not assume a linear relationship between summary statistics and parameters, as is the case for L_1- and L_2Boosting. Instead, LogitBoost fits a logistic regression model, which might be more appropriate. On the other hand, it requires choosing a discretization procedure $f(\cdot)$ to map $\boldsymbol{\alpha}_t \in \mathbb{R}$ to $y \in \{0, 1\}$ (see below). Since such a choice is arbitrary, it would be problematic to use the resulting fit (a linear combination on the logit-scale) directly as $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$. In practice, we instead assigned a candidate statistic $\mathbf{S}^{(j)}$ $(j = 1, \ldots, q)$ to $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ if the corresponding boosted coefficient $\hat{\lambda}^{(j)}$ (*cf.* Equation 8) was different from zero and omitted it otherwise. Therefore, compared to L_1- and L_2Boosting, the reduction in dimensionality was on average lower, but the strong assumption of a linear relationship between $\boldsymbol{\alpha}^{(k)}$ and $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ was avoided. Note that, in principle, nonlinear relationships may be fitted with the L_1- and L_2-loss, too (Friedman *et al.* 2000). In File S1 we provide explicit expressions for the population minimizers (Equation 6) and some more insight on the boosting algorithms under the three loss functions used here.

Partial least-squares regression: Recently, Wegmann *et al.* (2009) proposed to choose summary statistics in ABC via PLS regression (*e.g.*, Hastie *et al.* 2011 and references therein). PLS is related to principal component regression. But in addition to maximizing the variance of the predictors $\mathbf{X}$, at the same time, it maximizes the correlation of $\mathbf{X}$ with the response Y. Applied to the choice of summary statistics, it therefore not only decorrelates the summary statistics, but also chooses them according to their relation to $\boldsymbol{\alpha}$. Hastie *et al.* (2011) argue that the first aspect dominates over the latter, however. The number r of PLS components to keep is usually determined based on some cross-validation procedure (see below). In the context of ABC, the r components are multiplied by the corresponding statistics $\mathbf{S}^{(j)}$ $(j \leq r)$ to obtain $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ (Wegmann *et al.* 2009).

Global vs. local choice

We have so far suggested that $\mathbf{S}_{\boldsymbol{\alpha}}$ is close to sufficient for estimating $\boldsymbol{\alpha}$. This will hardly be the case in practice. By definition, the optimal choice of $\mathbf{S}_{\boldsymbol{\alpha}}$ then depends on the unknown true parameter value(s). Ideally, we therefore want to focus the choice of $\mathbf{S}_{\boldsymbol{\alpha}}$ on the neighborhood of the truth. The latter is not known in practice. As a workaround, we propose to use the n simulated pairs $\langle \mathbf{s}'_{t^*}, \boldsymbol{\alpha}'_{t^*} \rangle$ from step B3 in algorithm B and the observed summary statistics $\mathbf{s}$ to approximately establish this neighborhood as follows.

Local choice of summary statistics in B3:

1. Consider the n pairs $\langle \mathbf{s}'_{t^*}, \boldsymbol{\alpha}'_{t^*} \rangle$ $(t^* = 1, \ldots, n)$ from step B3 in algorithm B.
2. Mean center each component $\mathbf{s}'^{(j)}$ $(j = 1, \ldots, q)$ and scale it to have unit variance.
3. Rotate $\mathbf{s}'$ using principal component analysis (PCA).
4. Apply the scaling from steps 2 and 3 to the observed summary statistics $\mathbf{s}$.
5. Mean center the PCA-scaled summary statistics obtained in step 3, and scale them to have unit variance. Do the same for the PCA-scaled observed statistics obtained in step 4. Denote the results by $\dot{\mathbf{s}}'$ and $\dot{\mathbf{s}}$, respectively.
6. For each $t^* \in n$, compute the Euclidean distance $\delta_{t^*} = \|\dot{\mathbf{s}}'_{t^*} - \dot{\mathbf{s}}_{t^*}\|$.
7. Keep the n' pairs $\langle \mathbf{s}'_{t^{**}}, \boldsymbol{\alpha}'_{t^{**}} \rangle$ $(t^{**} = 1, \ldots, n')$ for which $\delta_{t^*} \leq z$, where z is some threshold.
8. Use the n' points accepted in step 7 as a training set to choose statistics $\mathbf{S}_{\boldsymbol{\alpha}}$ with the desired method.
9. Continue with step B4 in algorithm B.

In step 2 above, the original summary statistics are brought to the same scale. Otherwise, summary statistics with a high variance would on average contribute relatively more to the Euclidean distance than summary statistics with a low variance. However, whether a simulated data point is far from or close to the target ($\mathbf{s}$) in multidimensional space may depend not only on the distance along the dimension of each statistic, but also on the correlation among statistics. This can be accounted for by decorrelating the statistics, as is done by PCA in step 3. In combination with the Euclidean distance in step 6, the procedure above essentially uses the Mahalanobis distance as a metric (Mahalanobis 1936). Although we cannot prove the optimality of this approach, it seems to work well in our simulations. Note that in steps 8 and 9, the summary statistics are used on their original scale again. This is because we want our method for choosing parameter-specific combinations of statistics to use the information comprised in the difference in scale among the original statistics—even in the vicinity of $\mathbf{s}$. The PCA scaling in step 5 is only used temporarily to determine δ_{t^*} in step 6. Figure S2 visualizes the different scales and the effect of determining an approximate neighborhood around $\mathbf{s}$.

The scheme just described may be combined with any of the methods for choosing summary statistics described above. In our case, we considered ABC with global and local versions of PLS (called *pls.glob* and *pls.loc* in the following), LogitBoost (*lgb.glob* and *lgb.loc*), L_1Boosting (*l1b.glob* and *l1b.loc*), and L_2Boosting (*l2b.glob* and *l2b.loc*). Moreover, we performed ABC with all candidate statistics $\mathbf{S}$ (*all*) as a reference.

Candidate summary statistics: Our set $\mathbf{S}$ of candidate summary statistics consisted of the mean and standard devation across loci of the following statistics: the average within-deme variance of allele length, the average within-deme gene diversity (H_1), the average between-deme gene diversity (H_2), the total F_{IS}, the total F_{ST}, the total within-deme mean squared difference (MSD) in allele length (S_1), the total between-deme MSD in allele length (S_2), the total R_{ST}, and the number of allele types in the total population. This amounts to a total of 18 summary statistics. We computed H_1, H_2, F_{IS}, and F_{ST} according to Nei and Chesser (1983) and S_1, S_2, and R_{ST} according to Slatkin (1995). Note that all summary statistics are symmetrical with respect to the order of the loci, which is consistent with our hierarchical parameterization of the ancestral mutation rate.

Implementation: Throughout, we used the prior distributions given in Table 1. In algorithm B, we performed $N = 10^6$ simulations and in B2i we assumed that $\pi(\boldsymbol{\alpha}, \tilde{\mathbf{m}}) = \pi(\boldsymbol{\alpha})\pi(\tilde{\mathbf{m}})$. In B3, we used $n = 10^4$ simulations for the choice of summary statistics (in both the global and the local versions). Moreover, we first chose sets of summary statistics for each parameter separately and then took the union of the sets, *i.e.*, $\mathbf{S}_{\boldsymbol{\alpha}} = \cup_k \mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$, where each $\mathbf{S}_{\boldsymbol{\alpha}^{(k)}}$ is chosen according to one of the methods proposed. This also applies to step 8 in the procedure for the local choice of summary statistics (see above). For the local choice, we kept the $n' = 1000$ pairs closest to the observation $\mathbf{s}$, and we used the pcrcomp function in R version 2.11 (R Development Core Team 2011) for PCA. Note that the set of the n' simulations closest to $\mathbf{s}$ and, hence, z in step 7 of the procedure for the local choice were the same for all local methods compared. In B5, we mean centered the summary statistics and scaled them to have unit variance. In B6, we chose the Euclidean distance as metric $\rho(\cdot)$. In B7 we did post-rejection adjustment with a weighted local-linear regression with weights from an Epanechnikov kernel (Beaumont *et al.* 2002), without additional scaling of parameters. For steps B6 and B7 we used the abc package (Csilléry *et al.* 2011) for R. We estimated the parameters and performed the linear regression on the same scale as the respective priors were defined.

For the PLS method, we used the pls package (Mevik and Wehrens 2007) for R and followed Wegmann *et al.* (2009, 2010). Specifically, we performed a Box–Cox transformation of the summary statistics prior to the PLS regression, and we chose the number of components to keep based on a plot of the root mean squared prediction error. We kept $r = 10$ components, both for *pls.glob* and *pls.loc* (Figure S3). For all methods based on boosting, we mean centered the summary statistics before boosting and used the glmboost function of the mboost package (Bühlmann and Hothorn 2007; Hothorn *et al.* 2010) for R. For the LogitBoost methods, we chose for each k the first and third quartiles of the sample of $\boldsymbol{\alpha}^{(k)}$ drawn in step B3 of algorithm B3 as the centers of the two classes of parameter values. For *lgb.glob*, we then assigned the 500 $\boldsymbol{\alpha}^{(k)}$-values closest to the first quartile to the first class ($y = 0$) and the 500 values closest to the third quartile to the second class ($y = 1$). For lgb.loc, we analogously assigned the 100 $\boldsymbol{\alpha}^{(k)}$-values closest to the two quartiles to the two classes. For both *lgb.glob* and *lgb.loc*, we chose the optimal m_{stop} based on the AIC (Akaike 1974; Bühlmann and Hothorn 2007), but set an upper limit for m_{stop} of 500 iterations. For *l1b.glob* and *l1b.loc*, we chose m_{stop} via 10-fold cross-validation with the cvrisk function of the mboost package, setting an upper limit of 100. Finally, for *l2b.glob* and *l2b.loc*, we chose m_{stop} based on the AIC, with an upper limit of 100. Figure S4, Figure S5, and Figure S6 further illustrate the boosting procedure.

Simulation study and application to data

To assess the performance of the different methods for choosing summary statistics and to study the influence of the rejection tolerance ε, we carried out a simulation study. For each $\varepsilon \in \{0.001, 0.01, 0.1\}$, we simulated 500 test data sets with parameter values sampled from the prior distributions and then inferred the posterior distribution for each set. In the case of local choice of summary statistics, the procedure of defining informative summary statistics based on the candidate statistics was run for each test data set separately. For the global choice, it was run only once per method, because there is no dependence on the supposed true value. Similar to Wegmann *et al.* (2009), we used as a measure of accuracy of the marginal posterior distributions the root mean integrated squared error (RMISE), defined as $\text{RMISE}_k = \sqrt{\int_{\Phi^{(k)}} (\boldsymbol{\phi}^{(k)} - \mu_k)^2 \pi(\boldsymbol{\phi}^{(k)}|\mathbf{s}) d\boldsymbol{\phi}^{(k)}}$, where μ_k is the true value of the kth component of the parameter vector $\boldsymbol{\phi}$ and $\pi(\boldsymbol{\phi}^{(k)}|\mathbf{s})$ is the corresponding estimated marginal posterior density. Recall that $\boldsymbol{\phi} = \boldsymbol{\alpha} = (\mu_{\theta_{\text{anc}}}, \sigma_{\theta_{\text{anc}}}, \omega)$ in our case. From this, we obtained the relative absolute RMISE (RARMISE) as $\text{RARMISE}_k = \text{RMISE}_k/|\mu_k|$. We also computed the absolute error (AE_k) between three marginal posterior point estimates (mode, mean, and median) and μ_k. Dividing by $|\mu_k|$, we obtained the relative absolute error (RAE_k). To directly compare the various methods to ABC with all summary statistics, we computed standardized variants of the RMISE and AE as follows: If a_k^{all} is the measure of accuracy for ABC with all summary statistics, and a_k^* is the one for ABC with the method of interest, the standardized measure was obtained as a_k^*/a_k^{all}. Importantly, we also assessed whether—across the 500 test data sets—the values obtained by evaluating the cumulative posterior distribution function at the respective true parameter value were uniformly distributed in [0, 1]. This indicates whether an inferred posterior distribution has converged to a distribution with correct coverage properties, given the respective computational constraints and summary statistics. We refer to this criterion as "coverage property" or "uniform distribution of posterior probabilities." This approach has been motivated by Cook *et al.* (2006) and applied in previous ABC studies (*e.g.*, Wegmann *et al.* 2009). Note that Cook *et al.* (2006) called these posterior probabilities "posterior quantiles," which is somewhat misleading. We tested for a uniform distribution of the posterior probabilities, using a Kolmogorov–Smirnov test (Sokal and Rohlf 1981). Since 81 such tests had to be performed, it would at first glance seem appropriate to correct for multiple testing. However, we want to protect ourselves from keeping by mistake the null hypothesis of uniformly distributed posterior probabilities, rather than to avoid rejection of the null hypothesis in marginal cases. Therefore, correcting for multiple testing would be conservative in the wrong direction. As a measure of our skepticism against uniformly distributed posterior probabilities, we report the Kolmogorov–Smirnov distance

$$\text{KS}_n = \sup_x |F_n(x) - F(x)|, \qquad (13)$$

where $F_n(x)$ is the empirical distribution function of n identically and independently distributed observations x_i from a random variable X, and $F(x)$ is the null distribution function (the uniform distribution between 0 and 1 in our case).

For the application to Alpine ibex, we used allele frequency and repeat length data from 37 putatively neutral microsatellites as described in Biebach and Keller (2009) (Figure 1 and Table S1). The data were provided to us by the authors. ABC simulations and inference were identical to those in the simulation study, with the same number of markers (see also File S1). The program called SPoCS that we wrote and used for simulation of the ibex scenario and a collection of R and shell scripts used for inference are available on the website http://pub.ist.ac.at/~saeschbacher/phd_e-sources/.

Results

Comparison of methods for choice of summary statistics

We have suggested boosting with componentwise linear regression as a base procedure for choosing summary statistics in ABC. Three loss functions were considered: the L_1- and the L_2-loss and the negative binomial log-likelihood. We have compared the performance of ABC with summary statistics chosen via different types of boosting to that of ABC with statistics chosen via PLS (Wegmann *et al.* 2009) and to that of ABC with all candidate summary statistics (Table 2). The RAE behaved similarly for the three point estimates (mode, mean, and median), but the mode was less reliable in cases where the posterior distributions did not have a unique mode (Figure S7). We decided to focus on the median. For assessment of the methods, we sought a low RARMISE and a low RAE of the median (RAE$_{\text{median}}$ in the following), and we required that the distribution of posterior probabilities of the true value did not deviate from uniformity for any parameter.

ABC with all summary statistics (*all*) and ABC with LogitBoost (*lgb.glob*) performed well in terms of RARMISE and RAE$_{\text{median}}$, especially when estimating $\mu_{\theta_{\text{anc}}}$ and ω (Figure 3, A and B). However, the posteriors of $\mu_{\theta_{\text{anc}}}$ inferred with all and *lgb.glob* tended to be biased (Kolmogorov–Smirnov distance and coverage P-value in Table 2).

Figure S8 implies that *all* yielded too narrow a posterior on average (U-shaped distribution of posterior probabilities of the true value), while *lgb.glob* tended to underestimate $\mu_{\theta_{\text{anc}}}$ (left-skewed distribution of posterior probabilities). This made us disfavor the methods *all* and *lgb.glob*. Throughout, ABC with L_1- and L_2Boosting on the global scale (*l1b.glob* and *l2b.glob*) performed very similarly in terms of RARMISE and RAE$_{\text{median}}$ (Figure 3, A and B). Because the L_2-loss is in general more sensitive to outliers, similarity in performance of *l1b.glob* and *l2b.glob* suggests that there were no problems with outliers, *i.e.*, no simulations producing extreme combinations of parameters and summary statistics. The accuracy of the *pls.glob* method was intermediate, except for the RAE$_{\text{median}}$ of $\mu_{\theta_{\text{anc}}}$ and $\sigma_{\theta_{\text{anc}}}$, where *pls.glob* performed worst (Figure 3B). For all methods, the RARMISE and the RAE$_{\text{median}}$ were considerably lower for $\mu_{\theta_{\text{anc}}}$ than for $\sigma_{\theta_{\text{anc}}}$ and ω. This implies that the latter two are more difficult to estimate with the data and model given here (see Figure S7). For an idea of how the data drive the parameter estimates, it is instructive to consider the correlation of individual summary statistics with the parameters (see Figure S11, Figure S12, and Figure S13).

The accuracy of estimation is expected to depend on the acceptance rate ε in a way determined by a trade-off between bias and variance (*e.g.*, Beaumont *et al.* 2002). While the RAE measures only the error of the point estimator, the RARMISE is a joint measure of bias and variance across the whole posterior distribution. The variance may be assigned to different sources. A first component—call it *simulation variance*—is a consequence of the finite number N of simulations. The lower ε is, the fewer points are accepted in the rejection step (B6 of algorithm B, see above). Posterior densities estimated from fewer points will be less stable than those inferred from more points, *i.e.*, show higher variance around the true posterior. A second variance component—the *sampling variance*—is due to the loss of information caused by using summary statistics that are not sufficient. To illustrate the trade-off between simulation and sampling variance, assume ε is fixed. If a large number of summary statistics are chosen, these may extract most of the information and thus limit the sampling variance. However, more summary statistics mean more dimensions and therefore a lower chance of accepting the same number of simulations than with fewer summary statistics and hence a higher simulation variance. In addition, accepting with $\delta_\varepsilon > 0$—which is characteristic of ABC—will introduce a systematic bias if the multidimensional density is not symmetric on the chosen metric with respect to the observation $\mathbf{s}$. On the other hand, increasing δ_ε reduces the simulation variance. Hence, there are in fact multiple trade-offs. It is not obvious in advance which one will dominate, and it is hard to make a prediction. This is reflected in our results: We found no uniform pattern for the dependence on ε of the RARMISE and the RAE$_{\text{median}}$. For instance, with *l2b.glob* the RARMISE increased as a function of ε for $\sigma_{\theta_{\text{anc}}}$, but decreased for ω (Figure 3A). Moreover, and typically for a trade-off, the relationship between accuracy and ε need not be monotonic (Figure 3) (*cf.* Beaumont *et al.* 2002).

Attempting to mitigate the lack of sufficiency, we have proposed to choose summary statistics locally—in the putative neighborhood of the true parameter values—rather than globally over the whole prior range. As expected, the local choice led to different combinations of statistics, and it had an effect on the scaling of the statistics for *pls.loc*, *l1b.loc*, and *l2b.loc* (Figure S14). However, the local versions of the different methods performed similarly to their global counterparts in terms of RARMISE and RAE$_{\text{median}}$ (Table 3 and Figure 3). The only exception to this is PLS when estimating $\mu\{\_\{\theta\_anc\}\}$, where the local version (*pls.loc*) resulted in an estimation error that increased more strongly

Table 2 Accuracy of different methods for choosing summary statistics on a global scale

Method	ε	Parameter	RARMISE[a]	RAE[b] mode	RAE mean	RAE median	KS_{500}[c]	Cov. P[d]
all	0.001	$\mu_{\theta_{\text{anc}}}$	0.143 (0.147)	0.062 (0.074)	0.065 (0.075)	0.062 (0.075)	0.072	0.011*
		$\sigma_{\theta_{\text{anc}}}$	0.452 (0.231)	0.269 (0.213)	0.269 (0.222)	0.265 (0.218)	0.034	0.610
		ω	0.446 (0.272)	0.221 (0.225)	0.215 (0.218)	0.219 (0.22)	0.027	0.859
	0.01	$\mu_{\theta_{\text{anc}}}$	0.141 (0.145)	0.061 (0.072)	0.064 (0.074)	0.065 (0.075)	0.082	0.003*
		$\sigma_{\theta_{\text{anc}}}$	0.466 (0.257)	0.299 (0.21)	0.286 (0.225)	0.282 (0.226)	0.019	0.992
		ω	0.432 (0.259)	0.233 (0.232)	0.226 (0.23)	0.232 (0.232)	0.026	0.880
	0.1	$\mu_{\theta_{\text{anc}}}$	0.140 (0.134)	0.065 (0.075)	0.067 (0.078)	0.067 (0.075)	0.081	0.003*
		$\sigma_{\theta_{\text{anc}}}$	0.463 (0.272)	0.324 (0.238)	0.306 (0.248)	0.296 (0.243)	0.032	0.677
		ω	0.431 (0.263)	0.234 (0.229)	0.228 (0.22)	0.226 (0.223)	0.038	0.482
pls.glob	0.001	$\mu_{\theta_{\text{anc}}}$	0.171 (0.16)	0.077 (0.087)	0.083 (0.089)	0.081 (0.088)	0.038	0.466
		$\sigma_{\theta_{\text{anc}}}$	0.488 (0.276)	0.291 (0.223)	0.289 (0.252)	0.276 (0.228)	0.024	0.936
		ω	0.451 (0.275)	0.238 (0.221)	0.234 (0.224)	0.237 (0.227)	0.022	0.969
	0.01	$\mu_{\theta_{\text{anc}}}$	0.166 (0.152)	0.080 (0.09)	0.079 (0.09)	0.079 (0.089)	0.035	0.562
		$\sigma_{\theta_{\text{anc}}}$	0.480 (0.291)	0.307 (0.223)	0.295 (0.268)	0.293 (0.242)	0.038	0.473
		ω	0.441 (0.262)	0.241 (0.234)	0.230 (0.225)	0.229 (0.226)	0.035	0.562
	0.1	$\mu_{\theta_{\text{anc}}}$	0.171 (0.146)	0.083 (0.091)	0.086 (0.097)	0.087 (0.094)	0.037	0.497
		$\sigma_{\theta_{\text{anc}}}$	0.469 (0.283)	0.319 (0.237)	0.307 (0.286)	0.310 (0.276)	0.056	0.089
		ω	0.433 (0.265)	0.240 (0.226)	0.234 (0.224)	0.234 (0.23)	0.049	0.178
lgb.glob	0.001	$\mu_{\theta_{\text{anc}}}$	0.149 (0.152)	0.064 (0.074)	0.065 (0.076)	0.064 (0.074)	0.082	0.002*
		$\sigma_{\theta_{\text{anc}}}$	0.435 (0.204)	0.270 (0.231)	0.261 (0.214)	0.247 (0.205)	0.038	0.466
		ω	0.456 (0.275)	0.235 (0.23)	0.230 (0.237)	0.232 (0.224)	0.025	0.913
	0.01	$\mu_{\theta_{\text{anc}}}$	0.145 (0.15)	0.066 (0.076)	0.066 (0.078)	0.066 (0.076)	0.103	<0.001*
		$\sigma_{\theta_{\text{anc}}}$	0.450 (0.223)	0.281 (0.215)	0.269 (0.217)	0.258 (0.209)	0.046	0.238
		ω	0.436 (0.27)	0.235 (0.234)	0.222 (0.223)	0.225 (0.228)	0.025	0.916
	0.1	$\mu_{\theta_{\text{anc}}}$	0.147 (0.142)	0.068 (0.079)	0.067 (0.078)	0.069 (0.079)	0.135	<0.001*
		$\sigma_{\theta_{\text{anc}}}$	0.471 (0.284)	0.288 (0.209)	0.301 (0.249)	0.271 (0.233)	0.054	0.103
		ω	0.427 (0.259)	0.232 (0.222)	0.225 (0.216)	0.228 (0.22)	0.042	0.329
l1b.glob	0.001	$\mu_{\theta_{\text{anc}}}$	0.188 (0.178)	0.075 (0.087)	0.074 (0.087)	0.076 (0.088)	0.035	0.573
		$\sigma_{\theta_{\text{anc}}}$	0.445 (0.202)	0.271 (0.236)	0.261 (0.232)	0.256 (0.216)	0.023	0.954
		ω	0.487 (0.297)	0.251 (0.259)	0.226 (0.227)	0.232 (0.226)	0.031	0.723
	0.01	$\mu_{\theta_{\text{anc}}}$	0.178 (0.17)	0.075 (0.087)	0.075 (0.088)	0.075 (0.085)	0.031	0.711
		$\sigma_{\theta_{\text{anc}}}$	0.463 (0.217)	0.288 (0.24)	0.271 (0.238)	0.259 (0.221)	0.029	0.805
		ω	0.468 (0.288)	0.255 (0.262)	0.228 (0.222)	0.235 (0.233)	0.034	0.595
	0.1	$\mu_{\theta_{\text{anc}}}$	0.177 (0.173)	0.078 (0.092)	0.078 (0.094)	0.079 (0.094)	0.043	0.311
		$\sigma_{\theta_{\text{anc}}}$	0.508 (0.299)	0.307 (0.21)	0.304 (0.269)	0.290 (0.248)	0.051	0.144
		ω	0.449 (0.272)	0.238 (0.241)	0.237 (0.222)	0.239 (0.227)	0.031	0.716
l2b.glob	0.001	$\mu_{\theta_{\text{anc}}}$	0.183 (0.173)	0.075 (0.087)	0.074 (0.085)	0.074 (0.086)	0.029	0.794
		$\sigma_{\theta_{\text{anc}}}$	0.441 (0.202)	0.273 (0.229)	0.257 (0.228)	0.254 (0.212)	0.028	0.828
		ω	0.487 (0.296)	0.251 (0.257)	0.231 (0.226)	0.234 (0.229)	0.033	0.648
	0.01	$\mu_{\theta_{\text{anc}}}$	0.180 (0.173)	0.077 (0.087)	0.077 (0.088)	0.076 (0.087)	0.030	0.766
		$\sigma_{\theta_{\text{anc}}}$	0.459 (0.213)	0.278 (0.242)	0.262 (0.235)	0.259 (0.214)	0.028	0.815
		ω	0.470 (0.288)	0.253 (0.26)	0.231 (0.221)	0.237 (0.229)	0.037	0.497
	0.1	$\mu_{\theta_{\text{anc}}}$	0.176 (0.171)	0.080 (0.092)	0.080 (0.096)	0.080 (0.093)	0.041	0.365
		$\sigma_{\theta_{\text{anc}}}$	0.503 (0.281)	0.300 (0.213)	0.297 (0.249)	0.283 (0.253)	0.052	0.139
		ω	0.445 (0.267)	0.240 (0.24)	0.239 (0.227)	0.236 (0.225)	0.030	0.755

RARMISE and RAE are given as the median across 500 independent estimations with true values drawn from the prior (median absolute deviation in parentheses). $\sigma_{\theta_{\text{anc}}}$ and ω were estimated on the $\log_{10}$-scale. *$P < 0.05$ without correction for multiple testing; *cf.* Figure S8.

[a] Relative absolute root mean integrated squared error (see text) with respect to the true value.

[b] Relative absolute error with respect to the true value.

[c] Kolmogorov–Smirnov distance between empirical distribution of posterior probabilities of the true parameter and $U(0, 1)$.

[d] P-value from a Kolmogorov–Smirnov test.

with ε compared to the global version (*pls.glob*). More importantly, however, the coverage properties of the posteriors for $\mu_{\theta_{\text{anc}}}$ deteriorated for *pls.loc*, *l1b.loc*, and *l2b.loc* (Table 3), compared to their global versions (Table 2). The effect was weakest for *l2b.loc* and in general increased as a function of ε. The *pls.loc* method tended to overestimate $\mu_{\theta_{\text{anc}}}$, while *lgb.loc*, *l1b.loc*, and *l2b.loc* tended to underestimate it (Figure S9).

For direct comparison of methods, before averaging across test sets, we standardized the measures of accuracy relative to those obtained with all summary statistics (Figure 4). The only local method that, for all parameters, led to lower RARMISE and $\text{RAE}_{\text{median}}$ than its global version was *l2b.loc*. In contrast, *lgb.glob* and *lgb.loc* performed very similarly; *pls.loc* did worse than *pls.glob* for $\mu_{\theta_{\text{anc}}}$, but better than *pls.glob* for $\sigma_{\theta_{\text{anc}}}$ and ω. Overall, we chose *l2b.loc* with ε = 0.01 as our favored method. This configuration provided good coverage for all parameters (Table 3). At the same time, it had lower RARMISE and $\text{RAE}_{\text{median}}$ than *pls.glob*,

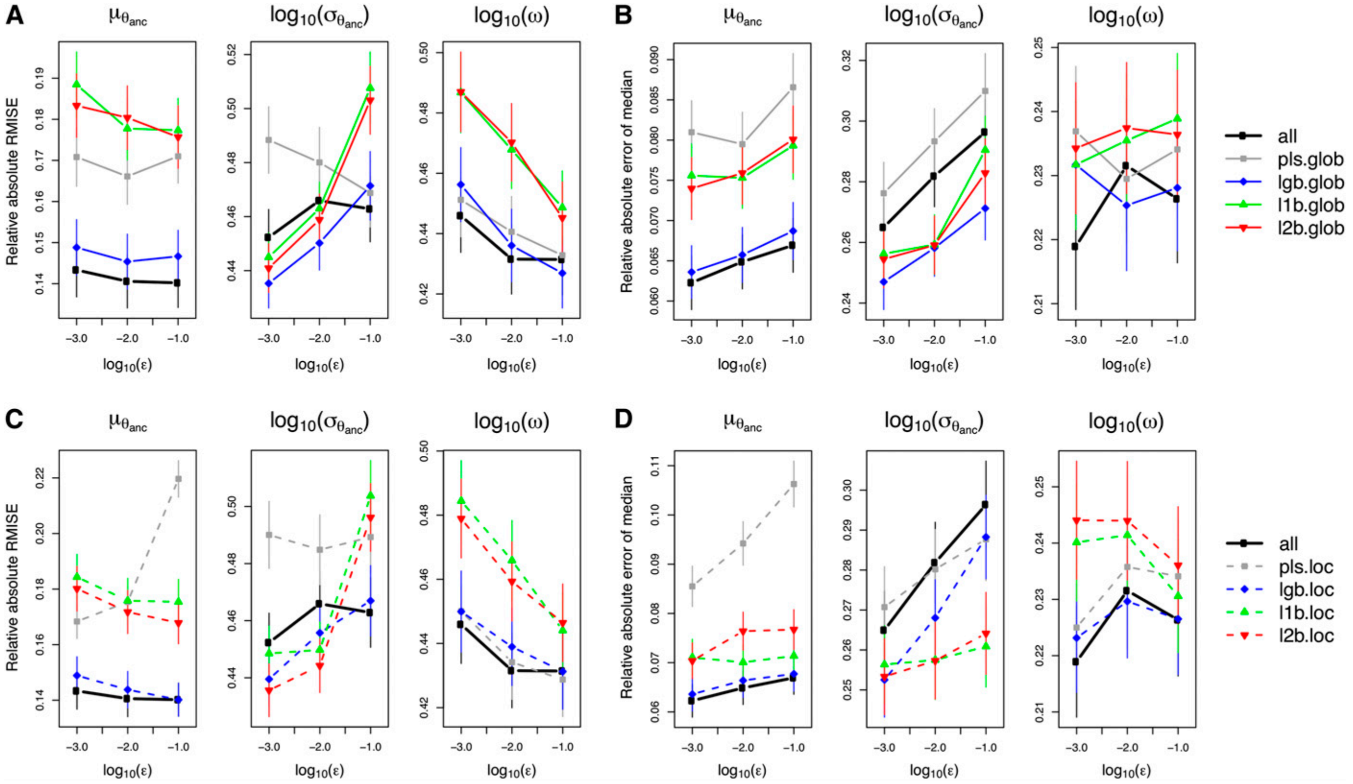


Figure 3 Accuracy of different methods for choosing summary statistics as a function of the acceptance rate (ε). (A and B) Results for different methods when applied to the whole parameter range (*global* choice). (C and D) The methods were applied only in the neighborhood of the (supposed) true value (*local* choice). The performance resulting from using all candidate summary statistics is shown for comparison in both rows. A and C show the root mean integrated squared error (RMISE), relative to the absolute true value. B and D give the absolute error of the posterior median, relative to the absolute true value. Plotted are the medians across $n = 500$ independent test estimations with true values drawn from the prior (error bars denote the median$\pm$MAD$/\sqrt{n}$, where MAD is the median absolute deviation).

the method that would also have had good coverage properties for $\mu_{\theta_{anc}}$. We disfavored all, *lgb.glob*, and *lgb.loc* due to their relatively weak coverage properties. Note that all methods compared in Figure 4 performed worse in terms of RARMISE and RAE$_{\text{median}}$ than *all* when estimating $\mu_{\theta_{anc}}$. This might be due to the loss of information caused by leaving out some summary statistics. Apparently, this loss is not fully compensated in our setting by the potential gain from reducing the dimensions. In models with many more dimensions, this may be different.

In summary, although performance in terms of RMISE and absolute error was only partially in favor of *l2b.loc*, we preferred this method based on its good coverage properties (Tables 2 and 3). Moreover, for $\log_{10}(\sigma_{\theta_{anc}})$ and $\log_{10}(\omega)$, the differences between methods measured by RMISE and absolute error were small compared to the error bars ($\pm\text{MAD}/\sqrt{n}$), implying that too much weight should not be given to the respective rankings in Figures 3 and 4.

It is worth recalling some of the characteristics of the methods compared here. The *pls* method is the only one that involves decorrelation of the statistics. Apparently, this did not lead to a net improvement compared to the other methods. Although one explanation might be that the statistics were only weakly correlated, Figure S10 shows evidence of strong correlation among some statistics. Thus, it would appear that correlation among statistics does not substantially reduce efficiency (but this finding cannot be readily extrapolated to other settings, as we have used only a moderate number of summary statistics here). The reduction of dimensions is strongest with the *l1b* and *l2b* methods, since they result in one linear predictor per parameter. On the other hand, these methods assume a linear relationship between parameters and statistics. Since the latter was clearly not the case (*e.g.*, Figure S11), it seems that the reduction of dimensions compensated for that assumption. This effect might be more pronounced in problems with many more statistics.

Application to Alpine ibex

Posterior distributions inferred for the ibex data with the various methods and $\varepsilon = 0.01$ are shown in Figure 5. The projection of some posterior density out of the prior support is not an artifact of kernel smoothing, but a consequence of regression adjustment. Leuenberger and Wegmann (2010) suggested a way of avoiding this problem. Since the effect is small—essentially absent for our favored method *l2b.loc*—we did not correct for this (*cf*. Figure S7). Moreover, the uniform distribution of posterior probabilities obtained with *l2b.loc* and $\varepsilon = 0.01$ (Figure S9) shows that the concerns

Table 3 Accuracy of different methods for choosing summary statistics on a local scale

Method	ε	Parameter	RARMISE		RAE mode		RAE mean		RAE median		KS_{500}	Cov. P
pls.loc	0.001	$\mu_{\theta_{\mathrm{anc}}}$	0.168	(0.136)	0.081	(0.091)	0.088	(0.095)	0.086	(0.091)	0.043	0.314
		$\sigma_{\theta_{\mathrm{anc}}}$	0.490	(0.262)	0.283	(0.229)	0.277	(0.234)	0.271	(0.226)	0.043	0.314
		ω	0.450	(0.278)	0.232	(0.234)	0.225	(0.228)	0.225	(0.228)	0.031	0.723
	0.01	$\mu_{\theta_{\mathrm{anc}}}$	0.175	(0.126)	0.088	(0.094)	0.098	(0.103)	0.094	(0.099)	0.067	0.023*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.485	(0.274)	0.287	(0.222)	0.287	(0.243)	0.280	(0.223)	0.046	0.232
		ω	0.434	(0.259)	0.240	(0.238)	0.235	(0.224)	0.236	(0.227)	0.033	0.655
	0.1	$\mu_{\theta_{\mathrm{anc}}}$	0.220	(0.147)	0.101	(0.103)	0.113	(0.108)	0.106	(0.104)	0.087	0.001*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.489	(0.282)	0.294	(0.216)	0.275	(0.243)	0.288	(0.231)	0.057	0.078
		ω	0.429	(0.259)	0.239	(0.226)	0.239	(0.227)	0.234	(0.223)	0.045	0.273
lgb.loc	0.001	$\mu_{\theta_{\mathrm{anc}}}$	0.149	(0.151)	0.061	(0.074)	0.067	(0.081)	0.064	(0.077)	0.076	0.006*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.440	(0.213)	0.271	(0.213)	0.259	(0.209)	0.253	(0.209)	0.037	0.500
		ω	0.450	(0.283)	0.229	(0.231)	0.223	(0.219)	0.223	(0.217)	0.029	0.794
	0.01	$\mu_{\theta_{\mathrm{anc}}}$	0.144	(0.147)	0.065	(0.074)	0.068	(0.078)	0.066	(0.077)	0.085	0.001*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.456	(0.237)	0.292	(0.209)	0.277	(0.223)	0.268	(0.213)	0.035	0.576
		ω	0.439	(0.27)	0.235	(0.229)	0.228	(0.225)	0.230	(0.225)	0.027	0.862
	0.1	$\mu_{\theta_{\mathrm{anc}}}$	0.140	(0.133)	0.068	(0.077)	0.069	(0.078)	0.068	(0.078)	0.093	<0.001*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.467	(0.275)	0.315	(0.233)	0.298	(0.24)	0.288	(0.234)	0.020	0.991
		ω	0.431	(0.264)	0.232	(0.22)	0.226	(0.219)	0.227	(0.222)	0.039	0.423
l1b.loc	0.001	$\mu_{\theta_{\mathrm{anc}}}$	0.184	(0.183)	0.070	(0.081)	0.070	(0.083)	0.071	(0.082)	0.059	0.062
		$\sigma_{\theta_{\mathrm{anc}}}$	0.449	(0.215)	0.263	(0.234)	0.254	(0.219)	0.256	(0.218)	0.034	0.610
		ω	0.484	(0.281)	0.246	(0.253)	0.232	(0.218)	0.240	(0.233)	0.034	0.610
	0.01	$\mu_{\theta_{\mathrm{anc}}}$	0.176	(0.18)	0.072	(0.081)	0.070	(0.083)	0.070	(0.082)	0.071	0.012*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.450	(0.218)	0.268	(0.25)	0.263	(0.23)	0.257	(0.221)	0.033	0.651
		ω	0.466	(0.279)	0.255	(0.265)	0.234	(0.22)	0.241	(0.234)	0.029	0.791
	0.1	$\mu_{\theta_{\mathrm{anc}}}$	0.175	(0.181)	0.076	(0.092)	0.072	(0.084)	0.071	(0.085)	0.107	<0.001*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.504	(0.276)	0.277	(0.234)	0.291	(0.251)	0.261	(0.227)	0.045	0.257
		ω	0.444	(0.267)	0.238	(0.236)	0.237	(0.227)	0.231	(0.225)	0.032	0.694
l2b.loc	0.001	$\mu_{\theta_{\mathrm{anc}}}$	0.180	(0.18)	0.071	(0.08)	0.074	(0.084)	0.070	(0.081)	0.043	0.314
		$\sigma_{\theta_{\mathrm{anc}}}$	0.436	(0.207)	0.249	(0.222)	0.251	(0.215)	0.253	(0.213)	0.030	0.759
		ω	0.479	(0.275)	0.257	(0.261)	0.233	(0.226)	0.244	(0.235)	0.037	0.500
	0.01	$\mu_{\theta_{\mathrm{anc}}}$	0.172	(0.173)	0.075	(0.085)	0.077	(0.087)	0.076	(0.087)	0.056	0.084
		$\sigma_{\theta_{\mathrm{anc}}}$	0.444	(0.211)	0.258	(0.246)	0.264	(0.225)	0.257	(0.215)	0.033	0.651
		ω	0.459	(0.276)	0.256	(0.276)	0.234	(0.228)	0.244	(0.236)	0.036	0.532
	0.1	$\mu_{\theta_{\mathrm{anc}}}$	0.168	(0.169)	0.077	(0.091)	0.076	(0.09)	0.077	(0.091)	0.128	<0.001*
		$\sigma_{\theta_{\mathrm{anc}}}$	0.496	(0.266)	0.277	(0.235)	0.289	(0.241)	0.264	(0.23)	0.044	0.284
		ω	0.446	(0.271)	0.239	(0.242)	0.237	(0.23)	0.236	(0.233)	0.035	0.579

Details are as in Table 2 (*cf.* Figure S9).

that motivate the approach by Leuenberger and Wegmann (2010) do not apply in our case. Point estimates and 95% highest posterior density (HPD) intervals obtained with *l2b.loc* are given in Table 4. Recall that $\mu_{\theta_{\mathrm{anc}}}$ and $\sigma_{\theta_{\mathrm{anc}}}$ are hyperparameters of the distribution of $\theta_{\mathrm{anc},l}$ across loci: $\log_{10}(\theta_{\mathrm{anc},l}) \sim N(\mu_{\theta_{\mathrm{anc}}}, \sigma^2_{\theta_{\mathrm{anc}}})$ (*cf.* Table 1). Inserting the estimates from Table 4, we obtained $\log_{10}(\theta_{\mathrm{anc},l}) \sim N(0.110,\ 0.163^2)$, which implies a mean $\hat{\theta}_{\mathrm{anc}}$ across loci of 1.288. The limits of the interval defined by $\hat{\mu}_{\theta_{\mathrm{anc}}} \pm 2\hat{\sigma}_{\theta_{\mathrm{anc}}}$ translate into (0.607, 2.735) on the scale of θ_{anc}. Remember that $\theta_{\mathrm{anc}} = 4N_{\mathrm{e}}u$; it measures the total genetic diversity present in the ancestral deme at time $t_1 = 1906$ (Figure 2), *i.e.*, at the start of the reintroduction phase. Although we were able to estimate θ_{anc} with relatively high precision, that does not immediately tell us about N_{e} or u without knowing one of the two. However, given some rough, independent estimates of N_{e} and u, we may assess whether our estimate $\hat{\theta}_{\mathrm{anc}} \approx 1.288$ is plausible. On the one hand, historical records of the census size of the ancestral Gran Paradiso deme are available. In combination with an estimate of the ratio of effective to census size, we may therefore obtain a rough estimate of N_{e}. Specifically, the census size of the Gran Paradiso deme (Figure 1) was estimated as <100 for the early 19th century (Stuwe and Nievergelt 1991; Scribner and Stuwe 1994), as 3000 for the early 20th century (Stuwe and Scribner 1989), and as 4000 for the year 1913 (Maudet *et al.* 2002). In addition, Scribner and Stuwe (1994) estimated for eight ibex demes in the Swiss Alps the effective population size from census estimates of the numbers of adult males and females. Their estimates of N_{e} were about one-third of the respective total census estimates. Together, these numbers suggest that a realistic range for the ancestral effective size N_{e} might be between 30 and 1300. On the other hand, estimates of the mutation rate u for microsatellites range from 10^{-4} to 10^{-2} per locus and generation (Di Rienzo *et al.* 1998; Estoup and Angers 1998). Combining these two ranges results in θ_{anc} ranging from $1.2 \times 10^{-2} \approx 10^{-2}$ to $5.2 \times 10 \approx 10^2$, suggesting that our estimate $\hat{\theta}_{\mathrm{anc}} \approx 1.288$ is plausible. Perhaps more interestingly, we may ask about the range *across loci* of u that is compatible with the range of $\hat{\theta}_{\mathrm{anc}}$ corresponding to $\hat{\mu}_{\theta_{\mathrm{anc}}} \pm 2\hat{\sigma}_{\theta_{\mathrm{anc}}}$ (0.607, 2.735). The underlying

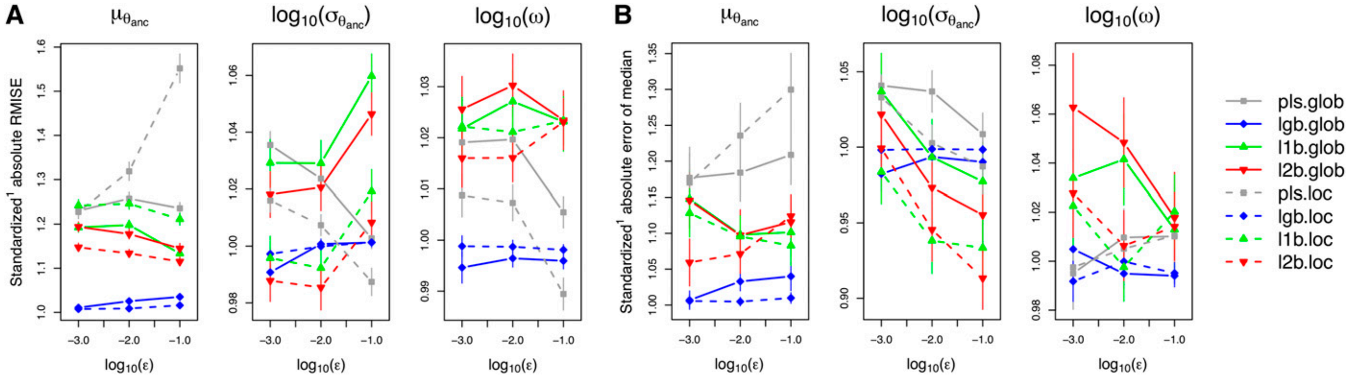


Figure 4 Standardized accuracy of different methods for choosing summary statistics as a function of the acceptance rate (ε). Standaridized^1 means that, before averaging across test sets, we divided the measures of accuracy for the respective method by the measure of accuracy obtained with all candidate summary statistics (this may change the relative order of methods compared to Figure 3, as the average of a ratio is generally not the same as the ratio of two averages). (A) Root mean integrated squared error (RMISE), relative to the RMISE obtained with all summary statistics. (B) Absolute error of the posterior median, relative to the one obtained with all summary statistics. Further details are as in Figure 3.

assumption is that N_e is roughly the same for all loci, so that variation in $\hat{\theta}_{anc}$ is exclusively due to variation of u across loci. Taking the geometric mean of the extremes from above, $\hat{N}_e = (30 \times 1300)^{1/2} \approx 197$, as a typical value, the corresponding interval for $\hat{u}$ across loci is $(7.7 \times 10^{-4}, 3.5 \times 10^{-3})$. In other words, most of the variation in u across loci spans less than one order of magnitude.

The estimates for $\log_{10}(\omega)$ from Table 4 imply a proportion of males obtaining access to matings of $\hat{\omega} \approx 0.208$ or ~21%. The 95% HPD interval for ω is (0.047, 0.934). An observational study in a free-ranging ibex deme suggested that ~10% of males reproduced (Aeschbacher 1978). More recently, Willisch *et al.* (2012) conducted a behavioral and genetic study and reported paternity scores for males of different age classes. The weighted mean across age classes from this study is ~14% successful males. Given the many factors that influence such estimates, our result of 21% seems in good agreement with these values, and our 95% HPD interval includes them. Two points are worth noting. First, our 95% HPD interval for ω seems large, which reflects the uncertainty involved in this parameter. Second, when estimating ω, we are essentially estimating the ratio of *recent* effective population size to census population size, $N_e^{(i)}/N$, where $N_e^{(i)}$ is the effective size of a derived deme d_i. This ratio may be smaller than one for many reasons—not just male mating access. Thus, we have strictly speaking estimated the strength of genetic drift due to deviations in reproduction from that in an idealized population. Nevertheless, the good agreement with the independent estimates of male mating access is striking.

In Figure 6, we report pairwise joint posterior distributions for *l2b.loc* and $\varepsilon = 0.01$. The pairwise joint modes are close to the marginal point estimates in Table 4. Moreover, Figure 6 suggests no strong correlation among parameters.

Discussion

We have suggested three variants of boosting for the choice of summary statistics in ABC and compared them to each other, to PLS regression, and to ABC with all candidate summary statistics. Moreover, we proposed to choose summary statistics locally, in the putative neighborhood of the observed data. Overall, the mean of the ancestral mutation rate $\mu_{\theta_{anc}}$ was more precisely estimated than its standard deviation $\sigma_{\theta_{anc}}$ and the male mating access rate ω. In our context, ABC with summary statistics chosen locally via boosting with componentwise linear regression as a base procedure and the L_2-loss performed best in terms of accuracy (measured by RARMISE and $\text{RAE}_{\text{median}}$) and uniformity of posterior probabilities together. However, the difference between the methods was moderate and the ranking depended to some degree on our choice of criteria to assess performance. If the main interest had been in a small error of point estimates (low $\text{RAE}_{\text{median}}$), but less in good overall posterior properties (low RARMISE and uniform posterior probabilities of the true value) at the same time, boosting with the negative binomial log-likelihood loss and, somewhat surprisingly, ABC with all candidate statistics, would have been preferable to boosting with the L_1- and L_2-loss. Under this criterion (low $\text{RAE}_{\text{median}}$), the performance of the PLS method was intermediate when estimating ω, but inferior to that of any boosting approach when estimating $\mu_{\theta_{anc}}$ and $\sigma_{\theta_{anc}}$. In general, choosing summary statistics locally slightly improved the accuracy compared to the global choice, but it led to worse posterior coverage for $\mu_{\theta_{anc}}$. The local version of L_2Boosting with acceptance rate $\varepsilon = 0.01$ coped best with this trade-off.

Applying that method to Alpine ibex data, we estimated the mean across loci of the scaled ancestral mutation rate as $\hat{\theta}_{anc} \approx 1.288$. The estimates for $\sigma_{\theta_{anc}}$ implied that most of the variation across loci of the mutation rate u was between 7.7×10^{-4} and 3.5×10^{-3}. The proportion of males obtaining access to matings per breeding season was estimated as $\hat{\omega} \approx 0.21$, which is in good agreement with recent independent estimates. This result suggests that the strong dominance hierarchy in Alpine ibex is reflected in overall genetic

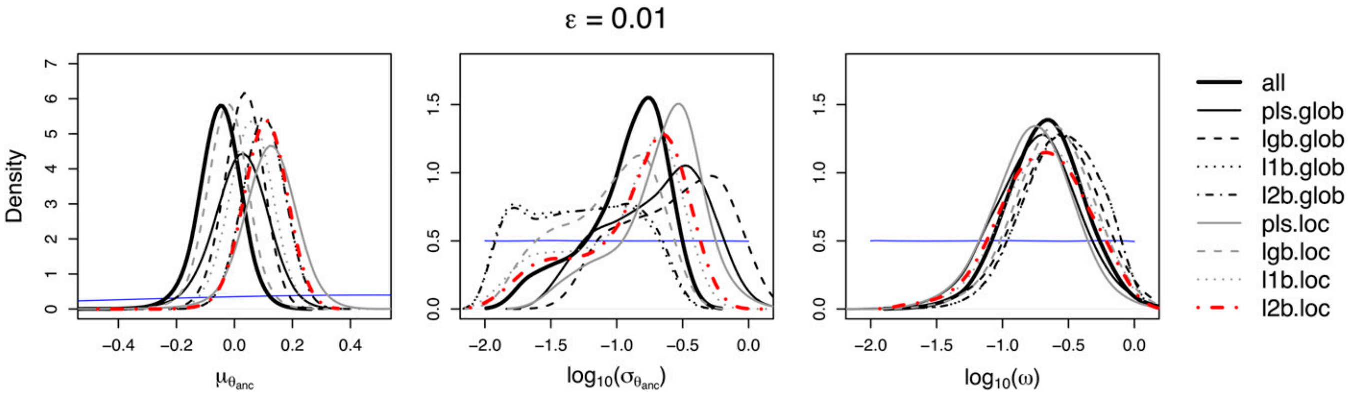


Figure 5 Marginal posterior distributions inferred from the Alpine ibex data. Posteriors obtained with tolerance ε = 0.01 and various methods for choosing summary statistics are compared. The dot-dashed red line corresponds to the method that performed best in the simulation study (*l2b.loc*; Tables 2 and 3 and Figures 3 and 4). Thin blue lines give the prior distribution (*cf.* Table 1). For pairwise joint posterior distributions, see Figure 6. Point estimates and 95% HPD intervals are given in Table 4.

diversity and should therefore be considered an important factor in determining the strength of genetic drift.

It should be noted that the results we reported here about the choice of summary statistics are specific to the model, to the data, and, in particular, to the choice of criteria used to assess performance. Another method may perform better under a different setting, and this is most likely a general feature of inference with ABC (*cf*. Blum *et al.* 2012). For the various points where some choice must be made—summary statistics, metric, algorithm, and postrejection adjustment—by nature, no single strategy is best in every case. Rather, the focus should be on choosing the best strategy for a specific problem. In practice, this implies comparing alternatives and assessing performance in a simulation study. Along these lines, there is still scope for new ideas concerning the various choices in ABC (see Beaumont *et al.* 2010). In particular, the choice of the metric makes ABC a scale-dependent method. This applies both to the ABC algorithm in general and to our suggestion of choosing summary statistics in the putative neighborhood of the truth. One could, for instance, use the Mahalanobis instead of the Euclidean distance, but even this is based on an assumption that is not necessarily appropriate (multivariate normal distribution of variables). In a specific application, one metric may do better than another, but it may not be obvious why. Overall, this poses an open problem and motivates future research (Wilkinson 2008).

As more data become available and more complex models are justifiable, it will be necessary that methods of inference keep pace. In principle, ABC is scalable and able to face this challenge. The problems arise in practice, and the combination of approaches devised to tackle them is itself becoming intricate. Researchers may be interested in a single program that implements these approaches and allows for inference with limited effort needed for tuning, simulation, and cross-validation. However, such software runs the risk of being treated as a black box. This problem is not unique to ABC, but equally applies to other sophisticated approaches of inference, such as coalescent-based genealogy samplers (Kuhner 2009). In the context of ABC, rather than having a single piece of software, we find it more promising to combine separate pieces of software that each implement a specific step. The appropriate combination must be chosen specifically for any application. It will always be necessary to evaluate the performance of any ABC method through simulation-based studies. Such a modular approach has recently been fostered by the developers of ABCtoolbox (Wegmann *et al.* 2010) or the abc package for R (Csilléry *et al.* 2011). Here, we contribute to this by providing a flexible simulation program that readily integrates into any ABC procedure.

Recently, two interesting alternative approaches have been proposed for choosing summary statistics with a focus on the putative location of the true parameter value, rather than the whole prior range. Nunes and Balding (2010) suggest a two-step procedure. Starting with a set of candidate summary statistics, at a first stage, standard ABC is carried out for (possibly) all subsets of these statistics, and the subset resulting in the posterior distribution with the minimum entropy is chosen. This subset of statistics is used to determine the n' simulations with the smallest Euclidean distance to the observation. At the second stage, the n' data sets close to the putative truth are used as a training set to choose, again, among (possibly) all subsets of the original candidate statistics. Here, Nunes and Balding (2010) propose as an optimization criterion the average square root of the sum of squared errors, averaged over the training data sets.

Fearnhead and Prangle (2012) follow the idea of optimizing the choice of summary statistics with respect to the accuracy of certain estimates of the parameters (*e.g.*, a point estimate), rather than the full posterior distribution. For instance, if the goal is to minimize the quadratic loss between the point estimate and the true value, the authors prove that the posterior mean is the optimal summary statistic. Since the posterior mean is not available in advance, they propose to first conduct a pilot ABC study to determine the region of high posterior mass. For this region, they then

Table 4 Posterior estimates for Alpine ibex data from ABC with summary statistics chosen locally via L_2Boosting and acceptance rate ε = 0.01

Parameter	Mode	Mean	Median	95% HPD[a] interval
$\mu_{\theta_{\text{anc}}}$	0.1089	0.1081	0.1101	(−0.0391, 0.2545)
$\log_{10}(\sigma_{\theta_{\text{anc}}})$	−0.6453	−0.8928	−0.7867	(−1.7615, −0.2613)
$\log_{10}(\omega)$	−0.6159	−0.6933	−0.6824	(−1.33, −0.0294)

[a] Highest posterior density.

draw parameters and simulate data to obtain training data sets. These are used in a third step to fit a linear regression with the parameters as responses and a vector-valued function of the original summary statistics as explanatory variables (allowing for nonlinear transformations of the original statistics). The linear fits are used as new summary statistics for the corresponding parameter. A final ABC run is then performed, with a prior restricted to the range established in the first step, and summary statistics chosen in the third step. Fearnhead and Prangle (2012) refer to this as *semi-automatic* and independent of the choice of statistics. However, as the authors note, it does depend on the initial choice of candidate statistics and on the choice of the vector-valued function. Moreover, if the (transposed) candidate statistics are uncorrelated, we suspect that their method would be equivalent to using the first component in a univariate PLS regression.

The approaches by Nunes and Balding (2010) and Fearnhead and Prangle (2012) and our local boosting procedures all consist of several steps, at least one being devoted to establishing the vicinity of the putative truth. While the method by Nunes and Balding (2010) and LogitBoost aim at choosing the "best subset" from a set of candidate statistics (without transforming them), the method by Fearnhead and Prangle (2012), PLS, and L_1- and L_2Boosting "construct" new summary statistics as functions of the original ones. The former has the advantage that the summary statistics conserve their interpretation, while the latter has the potential of better extracting and combining information contained partly in the various candidate statistics. The method by Nunes and Balding (2010) suffers from the fact that all subsets of candidate statistics must be explored, which is prohibitive in the case of large numbers of statistics. Here, boosting offers a potential advantage, because the functional gradient descent is a "greedy" algorithm (see *Appendix*). It does not explore *all* possible combinations of statistics, but in any iteration selects only one candidate statistic that improves an optimization criterion, given the current stage of the algorithm.

A direct comparison of all the recently proposed methods for the choice of statistics in ABC (*e.g.*, Joyce and Marjoram 2008; Wegmann *et al.* 2009; Nunes and Balding 2010; Jung and Marjoram 2011; Fearnhead and Prangle 2012) seems due. Nunes and Balding (2010) and Blum *et al.* (2012) compare a subset of these methods for a simple toy model with mutation and recombination in a panmictic population (*cf*. Joyce and Marjoram 2008). Blum *et al.* (2012) also include two examples from epidemiological modeling and material science. Their main conclusion is that the best method depends on the model. Crucial aspects seem to be the number of parameters, the number of candidate summary statistics, and the degree of collinearity of the statistics. Importantly, the PLS method (Wegmann *et al.* 2009)—although widely used in recent ABC applications—has been shown not to be very efficient and our results are consistent with this. However, a comparison for a range of relevant population genetic models, including some with larger numbers of parameters, is currently missing.

The boosting approach proposed here should also be suitable for constructing summary statistics to perform model comparison with ABC (*e.g.*, Fagundes *et al.* 2007; Blum and Jakobsson 2011). Despite recent criticism (Robert *et al.* 2011; but see Didelot *et al.* 2011), ABC-type model comparison remains an interesting option. By treating the model index as the single response, the algorithm proposed here might be used in this context.

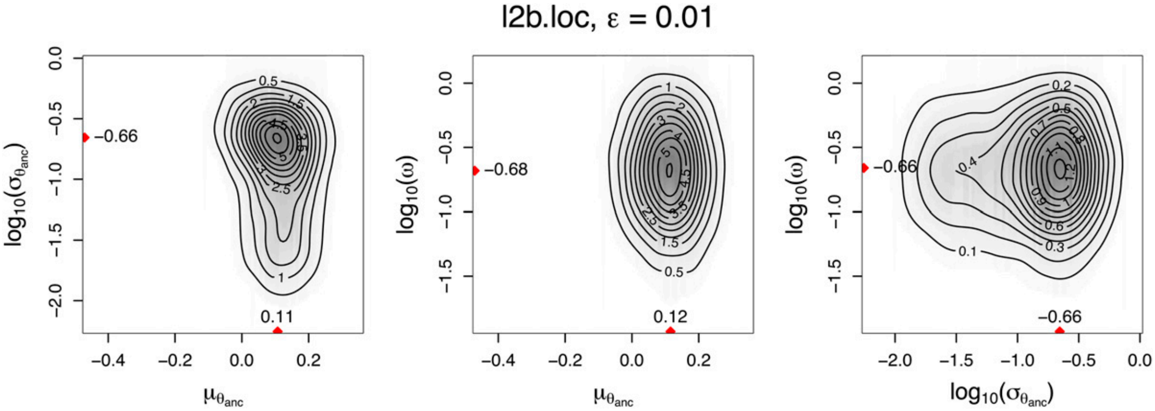


Figure 6 Pairwise joint posterior distributions given data observed in Alpine ibex, obtained with tolerance ε = 0.01 and summary statistics chosen locally via L_2Boosting (*l2b.loc*). Red triangles denote parameter values corresponding to the pairwise joint modes. Each time, the third parameter has been marginalized over.

Acknowledgments

We thank Nick Barton, Christoph Lampert, and three anonymous reviewers for helpful comments on earlier versions of the manuscript. We thank Iris Biebach and Lukas Keller for providing genetic data and for help in reconstructing the demographic data. We thank Markus Brülisauer, Erwin Eggenberger, Flurin Filli, Bernhard Nievergelt, Marc Rosset, Urs Zimmermann, Martin Zuber, the staff from Tierpark Langenberg, Tierpark Dählhölzli (Bern), Wildpark Peter and Paul (St. Gallen), Wildtier Schweiz, and the Swiss Federal Office for the Environment for providing information on population history and reintroduction, and Barbara Oberholzer for double checking the reintroduction history. We thank Walter Abderhalden, Iris Biebach, Michael Blum, Kati Csilléry, Lukas Keller, and Christian Willisch for discussion. This work made use of the computational resources provided by Institute of Science and Technology (IST) Austria and the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk). The ECDF is partially supported by the e-Science Data Information & Knowledge Transformation (eDIKT) initiative (http://www.edikt.org.uk). S.A. acknowledges financial support from IST Austria, the Janggen-Pöhn Foundation (St. Gallen), and the Roche Research Foundation (Basel) and from the University of Edinburgh in the form of a Torrance Studentship.

Literature Cited

Aeschbacher, A., 1978 *Das Brunftverhalten des Alpensteinbocks*. Eugen Rentsch Verlag, Erlenbach-Zürich, Switzerland.

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19: 716–723.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. B 355: 1553–1562.

Beaumont, M. A., 2010 Approximate Bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41: 379–406.

Beaumont, M. A., and B. Rannala, 2004 The Bayesian revolution in genetics. Nat. Rev. Genet. 5: 251–261.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin, and C. P. Robert, 2009 Adaptive approximate Bayesian computation. Biometrika 96: 983–990.

Beaumont, M. A., R. Nielsen, C. P. Robert, J. Hey, O. Gaggiotti *et al.*, 2010 In defence of model-based inference in phylogeography – reply. Mol. Ecol. 19: 436–446.

Bertorelle, G., A. Benazzo, and S. Mona, 2010 ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol. Ecol. 19: 2609–2625.

Biebach, I., and L. F. Keller, 2009 A strong genetic footprint of the re-introduction history of Alpine ibex (*Capra ibex ibex*). Mol. Ecol. 18: 5046–5058.

Blum, M., and O. François, 2010 Non-linear regression models for approximate Bayesian computation. Stat. Comput. 20: 63–73.

Blum, M. G. B., and M. Jakobsson, 2011 Deep divergences of human gene trees and models of human origins. Mol. Biol. Evol. 28: 889–898.

Blum, M. G. B., M. A. Nunes, D. Prangle, and S. A. Sisson, 2012 A comparative review of dimension reduction methods in approximate Bayesian computation. Stat. Sci. (in press).

Bühlmann, P., and T. Hothorn, 2007 Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. 22: 477–505.

Charlesworth, B., and D. Charlesworth, 2010 *Elements of Evolutionary Genetics*. Roberts & Company Publishers, Greenwood Village, Colorado.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Cook, S. R., A. Gelman, and D. B. Rubin, 2006 Validation of software for Bayesian models using posterior quantiles. J. Comput. Graph. Stat. 15: 675–692.

Couturier, M. A. J., 1962 *Alpine Ibex*. Chez l'auteur, Allier, France (in French).

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François, 2010 Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. 25: 410–418.

Csilléry, K., O. François, and M. G. B. Blum, 2012 Abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

Didelot, X., R. G. Everitt, A. M. Johansen, and D. J. Lawson, 2011 Likelihood-free estimation of model evidence. Bayesian Anal. 6: 49–76.

Diggle, P. J., 1979 On parameter estimation and goodness-of-fit testing for spatial point patterns. Biometrics 35: 87–101.

Diggle, P. J., and R. J. Gratton, 1984 Monte Carlo methods of inference for implicit statistical models. J. R. Stat. Soc. B 46: 193–227.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148: 1269–1284.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology*, Vol. 306, edited by G. R. Carvalho. IOS Press, Amsterdam.

Estoup, A., and J.-M. Cornuet, 1999 Microsatellite evolution: inference from population data, pp. 49–65 in *Microsatellites – Evolution and Application*, edited by D. B. Goldstein and C. Schloetterer. Oxford University Press, London/New York/Oxford.

Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA 104: 17614–17619.

Fearnhead, P., and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc. B 74: 419–474.

Fisher, R. A., 1922 On the mathematical foundations of theoretical statistics. Phil. Trans. R. Soc. A 222: 309–368.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, and L. L. e. Stuve, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Freund, Y., 1995 Boosting a weak learning algorithm by majority. Inform. Comput. 121: 256–285.

Freund, Y., and R. E. Schapire, 1996 Experiments with a new boosting algorithm, pp. 148–156 in *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann Publishers, San Francisco.

Freund, Y., and R. E. Schapire, 1999 A short introduction to boosting. J. Jpn. Soc. Artif. Intell. 14: 771–780.

Friedman, J., T. Hastie, and R. Tibshirani, 2000 Special invited paper. additive logistic regression: a statistical view of boosting. Ann. Stat. 28: 337–374.

Friedman, J. H., 2001 Greedy function approximation: a gradient boosting machine. Ann. Stat. 29: 1189–1232.

Fu, Y. X., and W. H. Li, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. 14: 195–199.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004 *Bayesian Data Analysis*, Ed. 2. Chapman & Hall/CRC, Boca Raton, Florida.

Ghosh, M., N. Reid, and D. A. S. Fraser, 2010 Ancillary statistics: a review. Stat. Sin. 20: 1309–1332.

Haldane, J. B. S., 1932 *The Causes of Evolution*, Ed. 2. Princeton University Press, Princeton, NJ.

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont *et al.*, 2005 Bayesian estimation of recent migration rates after a spatial expansion. Genetics 170: 409–417.

Hastie, T., R. Tibshirani, and J. Friedman, 2011 *The Elements of Statistical Learning – Data Mining*, *Inference*, *and Prediction*, Ed. 2. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner, 2010 Model-based boosting 2.0. J. Mach. Learn. Res. 11: 2109–2113.

Joyce, P., and P. Marjoram, 2008 Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. 7: 26.

Jung, H., and P. Marjoram, 2011 Choice of summary statistic weights in approximate Bayesian computation. Stat. Appl. Genet. Mol. Biol. 10: DOI:10.2202/1544-6115.1586.

Kuhner, M. K., 2009 Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24: 86–93.

Le Cam, L., 1964 Sufficiency and approximate sufficiency. Ann. Math. Stat. 35: 1419–1455.

Leuenberger, C., and D. Wegmann, 2010 Bayesian computation and model selection without likelihoods. Genetics 184: 243–252.

Lin, K., H. Li, C. Schlötterer, and A. Futschik, 2011 Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187: 229–244.

Mahalanobis, P. C., 1936 On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2: 49–55.

Marjoram, P., and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. 7: 759–770.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 100: 15324–15328.

Maudet, C., C. Miller, B. Bassano, C. Breitenmoser-Wursten, D. Gauthier *et al.*, 2002 Microsatellite DNA and recent statistical methods in wildlife conservation management: applications in Alpine ibex [*Capra ibex* (*ibex*)]. Mol. Ecol. 11: 421–436.

Maynard Smith, J., and J. Haigh, 1974 Hitch-hiking effect of a favorable gene. Genet. Res. 23: 23–35.

Mevik, B.-H., and R. Wehrens, 2007 The pls package: principal component and partial least squares regression in R. J. Stat. Softw. 18: 1–24.

Nei, M., and R. K. Chesser, 1983 Estimation of fixation indexes and gene diversities. Ann. Hum. Genet. 47: 253–259.

Nunes, M. A., and D. J. Balding, 2010 On optimal selection of summary statistics for approximate Bayesian computation. Stat. Appl. Genet. Mol. Biol. 9: 34.

Ohta, T., and M. Kimura, 1973 Model of mutation appropriate to estimate number of electrophoretically detectable alleles in a finite population. Genet. Res. 22: 201–204.

Pritchard, J., M. Seielstad, A. Perez-Lezaun, and M. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16: 1791–1798.

Raiffa, H., and R. Schlaifer, 1968 *Applied Statistical Decision Theory.* John Wiley & Sons, New York.

R Development Core Team, 2011 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Robert, C. P., J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, 2011 Lack of confidence in approximate Bayesian computation model choice. Proc. Natl. Acad. Sci. USA 108: 15112–15117.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298: 2381–2385.

Rubin, D. B., 1984 Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 12: 1151–1172.

Schapire, R. E., 1990 The strength of weak learnability. Mach. Learn. 5: 197–227.

Scribner, K. T., and M. Stuwe, 1994 Genetic relationships among Alpine ibex *Capra ibex* populations reestablished from a common ancestral source. Biol. Conserv. 69: 137–143.

Shao, J., 2003 *Mathematical Statistics*, Ed. 2. Springer-Verlag, New York.

Sisson, S. A., Y. Fan, and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 104: 1760–1765.

Sisson, S. A., Y. Fan, and M. M. Tanaka, 2009 Correction for Sisson *et al.*, Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 106: 16889.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457–462.

Sokal, R. R., and J. F. Rohlf, 1981 *Biometry – The Principles and Practice of Statistics in Biological Research*, Ed. 2. W. H. Freeman, New York.

Stuwe, M., and C. Grodinsky, 1987 Reproductive biology of captive Alpine ibex (*Capra i. ibex*). Zoo Biol. 6: 331–339.

Stuwe, M., and B. Nievergelt, 1991 Recovery of Alpine ibex from near extinction—the result of effective protection, captive breeding, and reintroduction. Appl. Anim. Behav. Sci. 29: 379–387.

Stuwe, M., and K. T. Scribner, 1989 Low genetic variablility in reintroduced Alpine ibex (*Capra ibex ibex*) populations. J. Mammal. 70: 370–373.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6: 187–202.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207–1218.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier, 2010 ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11: 116.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics 149: 1539–1546.

Wilkinson, R. D., 2008 Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355v1.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102: 7882–7887.

Willisch, C. S., and P. Neuhaus, 2009 Alternative mating tactics and their impact on survival in adult male Alpine ibex (*Capra ibex ibex*). J. Mammal. 90: 1421–1430.

Willisch, C., I. Biebach, U. Koller, T. Bucher, N. Marreros *et al.*, 2012 Male reproductive pattern in a polygynous ungulate with a slow life-history: the role of age, social status and alternative mating tactics. Evol. Ecol. 26: 187–206.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.

Communicating editor: N. A. Rosenberg

Appendix

Modular Inference for High-Dimensional Problems Using ABC

Here, we explore how ABC can be applied to complex situations, where a modular structure of the inferential problem can be exploited. For this purpose, we assume that the parameter vector $\boldsymbol{\phi}$ relevant to the problem can be split into two subvectors $\boldsymbol{\alpha}$ and $\tilde{\mathbf{m}}$ and that we have two corresponding vectors of summary statistics $\mathbf{S}_{\boldsymbol{\alpha}}$ and $\mathbf{S}_{\tilde{\mathbf{m}}}$, such that $\mathbf{S}_{\boldsymbol{\alpha}}$ contains most of the information on $\boldsymbol{\alpha}$, whereas $\mathbf{S}_{\tilde{\mathbf{m}}}$ contains most of the information on $\tilde{\mathbf{m}}$. It turns out that the modular structure can be exploited in such a situation, to split a high-dimensional problem into subproblems involving only lower-dimensional summary statistics.

To make this precise, we adapt the concepts of approximate sufficiency (*e.g.*, Le Cam 1964) and approximate ancillarity (Ghosh *et al.* 2010 and references therein). In the context of ABC, Joyce and Marjoram (2008) proposed an approach for choosing summary statistics based on approximate sufficiency.

In particular, we call $\mathbf{S}_{\boldsymbol{\alpha}}$ to be ε-sufficient for $\boldsymbol{\alpha}$ with respect to $\mathbf{S}_{\tilde{\mathbf{m}}}$, if

$$\sup_{\boldsymbol{\alpha}} \ln \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) - \inf_{\boldsymbol{\alpha}} \ln \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \leq \varepsilon \tag{A1}$$

for all $\tilde{\mathbf{m}}$. We further define $\mathbf{S}_{\boldsymbol{\alpha}}$ to be δ-ancillary with respect to $\tilde{\mathbf{m}}$, if

$$\sup_{\tilde{\mathbf{m}}} \ln \pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) - \inf_{\tilde{\mathbf{m}}} \ln \pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) < \delta \tag{A2}$$

for all $\boldsymbol{\alpha}$. Analogously, we define ε-sufficiency and δ-ancillarity for $\mathbf{S}_{\tilde{\mathbf{m}}}$ (note that ε and δ do not have the same meaning here as in the main text).

We first assume that $\mathbf{S}_{\boldsymbol{\alpha}}$ is ε-sufficient for $\boldsymbol{\alpha}$ relative to $\mathbf{S}_{\tilde{\mathbf{m}}}$ and δ-ancillary with respect to $\tilde{\mathbf{m}}$. Then,

$$\begin{aligned}
\pi(\boldsymbol{\alpha}|\mathbf{S}) &= \int \pi\left(\tilde{\mathbf{m}}, \boldsymbol{\alpha}|\mathbf{S}\right) d\tilde{\mathbf{m}} \\
&= \frac{\int \pi\left(\mathbf{S}_{\boldsymbol{\alpha}}, \mathbf{S}_{\tilde{\mathbf{m}}} \middle| \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi(\boldsymbol{\alpha}) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}}{\iint \pi\left(\mathbf{S}_{\boldsymbol{\alpha}}, \mathbf{S}_{\tilde{\mathbf{m}}} \middle| \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi(\boldsymbol{\alpha}) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}\, d\boldsymbol{\alpha}} \\
&= \frac{\int \pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi(\boldsymbol{\alpha}) \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}}{\iint \pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi(\boldsymbol{\alpha}) \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}\, d\boldsymbol{\alpha}} \\
&\leq \frac{\pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \boldsymbol{\alpha}\right) \pi\left(\boldsymbol{\alpha}\right) e^{\delta} \int \sup_{\boldsymbol{\alpha}} \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}}{e^{-\delta} \int \pi\left(\mathbf{S}_{\boldsymbol{\alpha}} | \boldsymbol{\alpha}\right) \pi\left(\boldsymbol{\alpha}\right) d\boldsymbol{\alpha} \int \inf_{\alpha} \pi\left(\mathbf{S}_{\tilde{\mathbf{m}}} \middle| \mathbf{S}_{\boldsymbol{\alpha}}, \tilde{\mathbf{m}}, \boldsymbol{\alpha}\right) \pi\left(\tilde{\mathbf{m}}\right) d\tilde{\mathbf{m}}} \\
&\leq \pi\left(\boldsymbol{\alpha} | \mathbf{S}_{\boldsymbol{\alpha}}\right) e^{2\delta+\epsilon}.
\end{aligned} \tag{A3}$$

A lower bound can be obtained in an analogous way, and we get

$$\pi(\boldsymbol{\alpha}|\mathbf{S}_{\boldsymbol{\alpha}}) e^{-2\delta-\epsilon} \leq \pi(\boldsymbol{\alpha}|\mathbf{S}) \leq \pi(\boldsymbol{\alpha}|\mathbf{S}_{\boldsymbol{\alpha}}) e^{2\delta+\epsilon}. \tag{A4}$$

If δ and ε are both small, a good approximation to the ABC-posterior $\pi(\boldsymbol{\alpha}\,|\,\mathbf{S})$ can therefore be obtained by using only $\mathbf{S}_{\boldsymbol{\alpha}}$.

Next, we look at the ABC posterior for $\tilde{\mathbf{m}}$ given $\boldsymbol{\alpha}$, $\pi(\tilde{\mathbf{m}}|\boldsymbol{\alpha}, \mathbf{S}_{\tilde{\mathbf{m}}})$, and start with Equation 4,

$$\pi\left(\tilde{\mathbf{m}}, \boldsymbol{\alpha}|\mathbf{S}\right) = \pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha}, \mathbf{S}\right) \pi\left(\boldsymbol{\alpha}|\mathbf{S}\right), \tag{A5}$$

and Equation 5,

$$\pi(\boldsymbol{\alpha}|\mathbf{S}) = \int \pi\left(\tilde{\mathbf{m}}, \boldsymbol{\alpha}|\mathbf{S}\right) d\tilde{\mathbf{m}}, \tag{A6}$$

from the main text, replacing the full data D by the summary statistics $\mathbf{S}$.

From (A6) it follows that a sample from the marginal posterior of $\boldsymbol{\alpha}$ can be obtained by taking the $\boldsymbol{\alpha}$-components of a sample from the joint posterior of $\tilde{\mathbf{m}}$ and $\boldsymbol{\alpha}$.

As shown above, $\pi(\boldsymbol{\alpha}\,|\,\mathbf{S})$ can be replaced without much loss by $\pi(\boldsymbol{\alpha}\,|\,\mathbf{S}_{\alpha})$, if $\mathbf{S}_{\tilde{\mathbf{m}}}$ is not informative for $\boldsymbol{\alpha}$ and $\mathbf{S}_{\boldsymbol{\alpha}}$ is not informative for $\tilde{\mathbf{m}}$. We show that $\pi(\tilde{\mathbf{m}}|\boldsymbol{\alpha},\mathbf{S}) = \pi(\tilde{\mathbf{m}}|\boldsymbol{\alpha},\mathbf{S}_{\boldsymbol{\alpha}},\mathbf{S}_{\tilde{\mathbf{m}}}) \approx \pi(\tilde{\mathbf{m}}|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}})$, given that $\mathbf{S}_{\tilde{\mathbf{m}}}$ is ε-sufficient for $\tilde{\mathbf{m}}$:

$$
\begin{aligned}
\pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\boldsymbol{\alpha}},\mathbf{S}_{\tilde{\mathbf{m}}}\right) &= \pi\left(\mathbf{S}_{\boldsymbol{\alpha}}\middle|\boldsymbol{\alpha},\tilde{\mathbf{m}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)\frac{\pi\left(\boldsymbol{\alpha},\tilde{\mathbf{m}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)}{\pi\left(\boldsymbol{\alpha},\mathbf{S}_{\boldsymbol{\alpha}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)} \\
&\le e^{\varepsilon}\pi\left(\mathbf{S}_{\boldsymbol{\alpha}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)\frac{\pi\left(\boldsymbol{\alpha},\tilde{\mathbf{m}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)}{\pi\left(\boldsymbol{\alpha},\mathbf{S}_{\boldsymbol{\alpha}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)} \\
&= e^{\epsilon}\pi\left(\mathbf{S}_{\boldsymbol{\alpha}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)\frac{\pi\left(\boldsymbol{\alpha},\tilde{\mathbf{m}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)}{\pi\left(\mathbf{S}_{\boldsymbol{\alpha}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)\pi\left(\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)} \\
&= e^{\epsilon}\frac{\pi\left(\boldsymbol{\alpha},\tilde{\mathbf{m}},\mathbf{S}_{\tilde{\mathbf{m}}}\right)}{\pi\left(\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)} \\
&= e^{\epsilon}\pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right).
\end{aligned}
\tag{A7}
$$

Together with an analogously obtained lower bound, we have that

$$
e^{-\varepsilon}\pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right) \le \pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\boldsymbol{\alpha}},\mathbf{S}_{\tilde{\mathbf{m}}}\right) \le e^{\varepsilon}\pi\left(\tilde{\mathbf{m}}\middle|\boldsymbol{\alpha},\mathbf{S}_{\tilde{\mathbf{m}}}\right)
\tag{A8}
$$

and again $\mathbf{S}_{\boldsymbol{\alpha}}$ can be omitted without much loss, if $\mathbf{S}_{\boldsymbol{\alpha}}$ does not provide much further information about $\tilde{\mathbf{m}}$ given $\mathbf{S}_{\tilde{\mathbf{m}}}$.

To summarize, breaking up ABC into lower-dimensional modules with separate summary statistics can be shown to lead to good approximations, if $\mathbf{S}_{\boldsymbol{\alpha}}$ and $\mathbf{S}_{\tilde{\mathbf{m}}}$ are ε-sufficient with respect to each other for their respective parameters. Also, $\mathbf{S}_{\boldsymbol{\alpha}}$ should be δ-ancillary for $\tilde{\mathbf{m}}$.

Functional Gradient Descent Boosting Algorithm

The general FGD algorithm for boosting, as given by Friedman (2001) and modified by Bühlmann and Hothorn (2007), is as follows.

FGD algorithm

1. Initialize $\hat{F}^{[0]}(\cdot) \equiv \arg\min_c n^{-1}\sum_{i=1}^{n} L(Y_i, c)$, set $m = 0$.
2. Increase m by 1. Compute the negative gradient and evaluate at $\hat{F}^{[m-1]}(\mathbf{X}_i)$:

$$
U_i = -\frac{\partial}{\partial F}L(Y_i, F)\big|_{F=\hat{F}^{[m-1]}(\mathbf{X}_i)}.
$$

3. Fit the negative gradient vector $(U_1, \ldots, U_n)$ to $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ by the base procedure:

$$
(\mathbf{X}_i, U_i)_{i=1}^{n} \rightarrow \hat{g}^{[m]}.
$$

4. Update $\hat{F}^{[m]}(\cdot) = \hat{F}^{[m-1]}(\cdot) + \nu\hat{g}^{[m]}(\cdot)$, where ν is a step-length factor.
5. Iterate steps 2–4 until $m = m_{\text{stop}}$.

Here, ν and m_{stop} are tuning parameters discussed in the main text. The result of this algorithm is a linear combination $\hat{F}(\cdot)$ of base procedure estimates, as shown in Equation 7 of the main text. In any specific version of boosting, the form of the initial function $\hat{F}^{[0]}(\cdot)$ in step 1 and the negative gradient U_i in step 2 may be expressed explicitly according to the loss function $L(\cdot,\cdot)$ (see File S1).

Base Procedure: Componentwise Linear Regression

We write the jth component of a vector v as $v^{(j)}$. The following base procedure performs simple componentwise linear regression,

$$
\begin{aligned}
\hat{g}(\mathbf{X}) &= \hat{\lambda}^{(\hat{\zeta})}\mathbf{X}^{(\hat{\zeta})},\\
\hat{\lambda}^{(j)} &= \frac{\sum_{i=1}^{n}\mathbf{X}^{(j)}U_i}{\sum_{i=1}^{n}\left(\mathbf{X}_i^{(j)}\right)^2},\\
\hat{\zeta} &= \arg\,\min_{1\le j\le p}\sum_{i=1}^{n}\left(U_i-\hat{\lambda}^{(j)}\mathbf{X}_i^{(j)}\right)^2,
\end{aligned}
\tag{A9}
$$

where $\hat{g}(\cdot)$, $\mathbf{X}$, and U_i are as in the FGD algorithm above. This base procedure selects the best variable in a simple linear model in the sense of ordinary least-squares fitting (Bühlmann and Hothorn 2007). To see this, note that $\hat{\lambda}^{(j)}$ in (A9) is the ordinary least-squares solution of a linear regression $U_i = \mathbf{X}_i^{(j)}\lambda^{(j)}$, in matrix form $\hat{\lambda}^{(j)} = (\mathbf{X}_i^{(j)\top}\mathbf{X}_i^{(j)})^{-1}\mathbf{X}_i^{(j)\top}U_i$. The choice of the loss functions enters indirectly via U_i (see File S1).

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.143164/-/DC1

A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation

Simon Aeschbacher, Mark A. Beaumont, and Andreas Futschik

DOI: 10.1534/genetics.112.143164

A novel approach for choosing summary statistics in approximate Bayesian computation

– Online supporting information –

Simon Aeschbacher[*,§], Mark A. Beaumont[**], Andreas Futschik[§§]

August 24, 2012

[*]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, [§]IST Austria (Institute of Science and Technology Austria), 3400 Klosterneuburg, Austria, [**]Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TW, United Kingdom, and [§§]Institute of Statistics and Decision Support Systems, University of Vienna, 1010 Vienna, Austria

SI Supporting Information: Overview

This file contains

- Additional Tables (Table S1)
- Additional Figures (Figures S1–S15)
- Additional Methods
- URLs to two Supporting Files (Files S2 and S3)

SI Supporting Information: Additional Tables

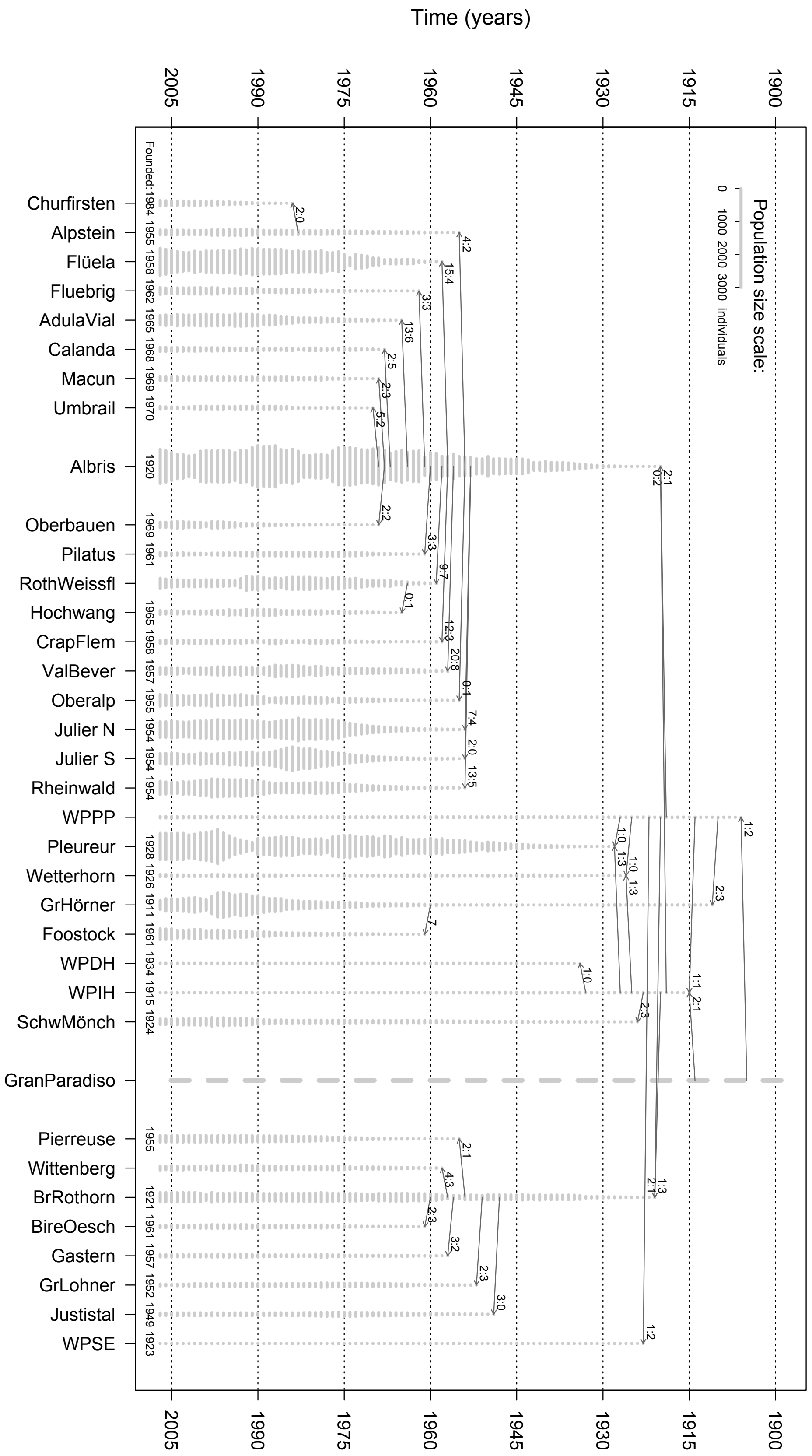
Table S1 Deme names, deme numbers and sampling sizes in the Alpine ibex data set

				Genetic sample size[c]		
Deme name	Deme number[a]	Short name	Internal number[b]	Males	Females	Total
Adula Vial	1	AdulaVial	100	21	16	37
Albris	2	Albris	101	28	33	61
Alpstein	3	Alpstein	102	12	18	30
Bire-Oeschinen	4	BireOesch	103	16	2	18
Brienzer Rothorn	5	BrRothorn	104	21	18	39
Calanda	6	Calanda	105	15	16	31
Churfirsten	7	Churfirsten	106	11	13	24
Crap da Flem	8	CrapFlem	107	16	11	27
Fluebrig	9	Fluebrig	108	17	15	32
Flüela	10	Flüela	109	37	38	75
Foostock	11	Foostock	110	9	18	27
Gastern	12	Gastern	111	5	6	11
Graue Hörner	13	GrHörner	112	21	26	47
Gross Lohner	14	GrLohner	113	15	7	22
Hochwang	15	Hochwang	114	14	14	28
Julier Nord	16	Julier N	115	12	11	23
Julier Süd	17	Julier S	116	12	11	23
Justistal	18	Justistal	117	15	4	19
Macun	19	Macun	118	12	10	22
Oberalp-Frisal	20	Oberalp	134	25	19	44
Oberbauenstock	21	Oberbauen	119	18	12	30
Pilatus	22	Pilatus	120	15	2	17
Mont Pleureur	23	Pleureur	121	22	7	29
Safien-Rheinwald	24	Rheinwald	122	22	13	35
Rothorn-Weissfluh	25	RothWeissfl	123	16	13	29
Schwarzmönch	26	SchwMönch	124	15	17	32
Umbrail	27	Umbrail	125	15	14	29
Val Bever	28	ValBever	126	20	12	32
Wetterhorn	29	Wetterhorn	127	9	10	19
Wittenberg	30	Wittenberg	128	15	6	21
Pierreuse-Gummfluh	31	Pierreuse	133	20	21	41
Wildpark Dählhölzli	32	WPDH	129	0	0	0
Wildpark Interlaken	33	WPIH	130	0	0	0
Wildpark St. Gallen	34	WPPP	131	0	0	0
Wildpark Seiler	35	WPSE	132	0	0	0

[a] As used in main text and Figure 1.
[b] As used in scripts and Supporting Files S2 and S3.
[c] The number of individuals from which genetic samples were taken, both in reality and in the simulations.

SI Supporting Information: Additional Figures

Figure S1 ***(facing page)*** Genealogy and demography of Alpine ibex demes analyzed in this study. Time goes from top to bottom, starting in the year 1900 and ending in 2007. Horizontal gray bars represent the known census sizes (Supporting File S2 *census sizes*) and arrows show the founder events by which demes were established. The numbers of males and females transferred are given close to the arrow head (males:females; for *Foostock*, the sex of the founders is unknown and only the total number of founders is given). Most demes received further individuals after the initial founder event, but these numbers are not shown here (see Supporting File S3 *transfers*). The deme ancestral to all other demes, *GranParadiso*, is shown as a vertical dashed line; its deme size is not known. See also Table S1 for the full deme names and Figure 1 for the geographical location of demes.


Time (years)
2005
1990
1975
1960
1945
1930
1915
1900
Population size scale:
0 1000 2000 3000 individuals
Founded:
Churfirsten 1984
Alpstein 1955
Flüela 1958
Fluebrig 1962
AdulaVial 1965
Calanda 1968
Macun 1969
Umbrail 1970
Albris 1920
Oberbauen 1969
Pilatus 1961
RothWeissfl
Hochwang 1965
CrapFlem 1958
ValBever 1957
Oberalp 1955
Julier N 1954
Julier S 1954
Rheinwald 1954
WPPP
Pleureur 1928
Wetterhorn 1926
GrHörner 1911
Foostock 1961
WPDH 1934
WPIH 1915
SchwMönch 1924
GranParadiso
Pierreuse 1955
Wittenberg
BrRothorn 1921
BireOesch 1961
Gastern 1957
GrLohner 1952
Justistal 1949
WPSE 1923
2:0
4:2
15:4
3:3
13:6
2:5
2:3
5:2
2:1
0:2
2:2
3:3
9:7
0:1
12:3
20:8
0:1
7:4
2:0
13:5
1:2
1:0
1:3
1:0
1:3
2:3
7
1:0
1:1
2:1
2:3
2:1
4:3
1:3
2:1
2:3
3:2
2:3
3:0
1:2

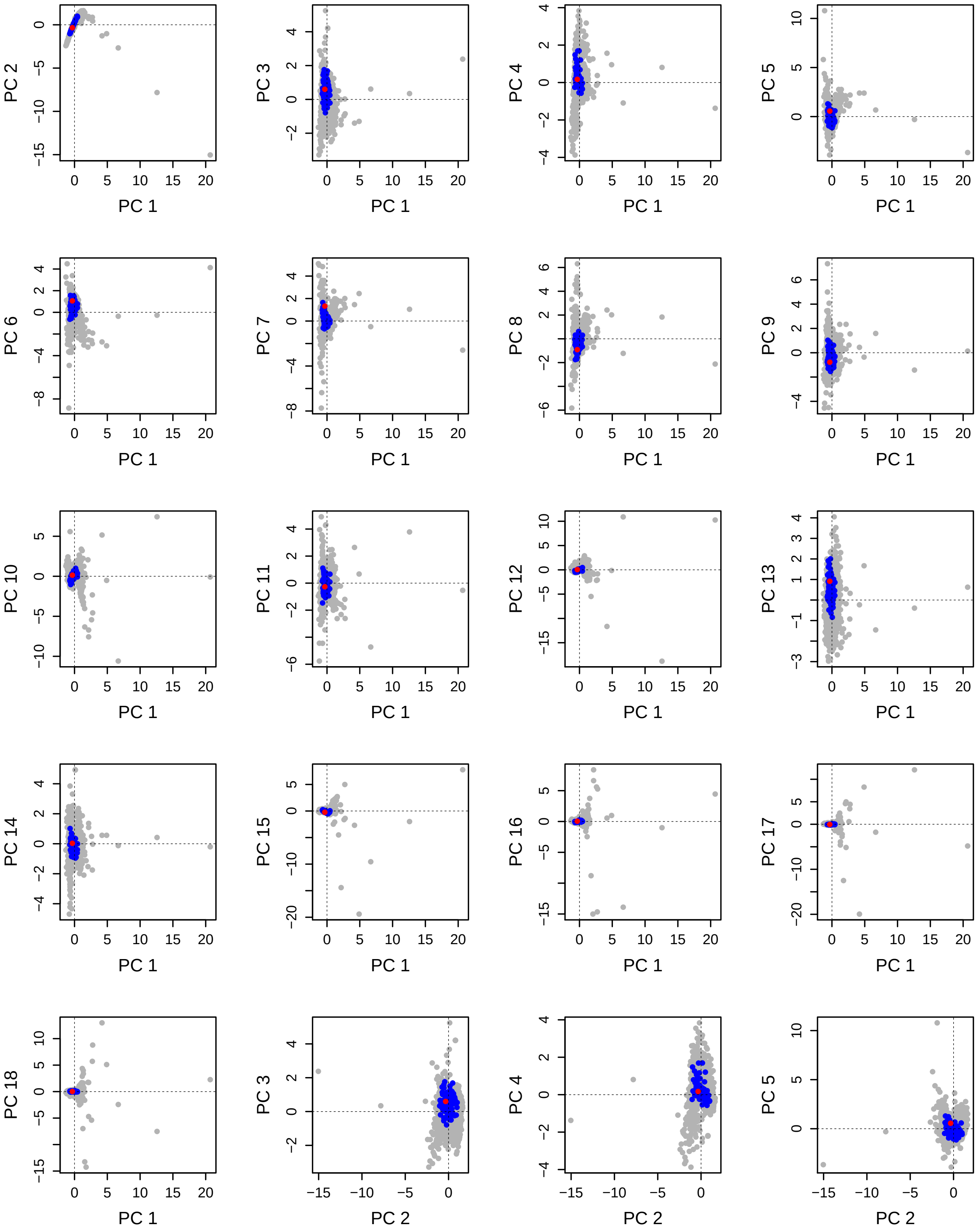


Figure S2 *Continued on next page*

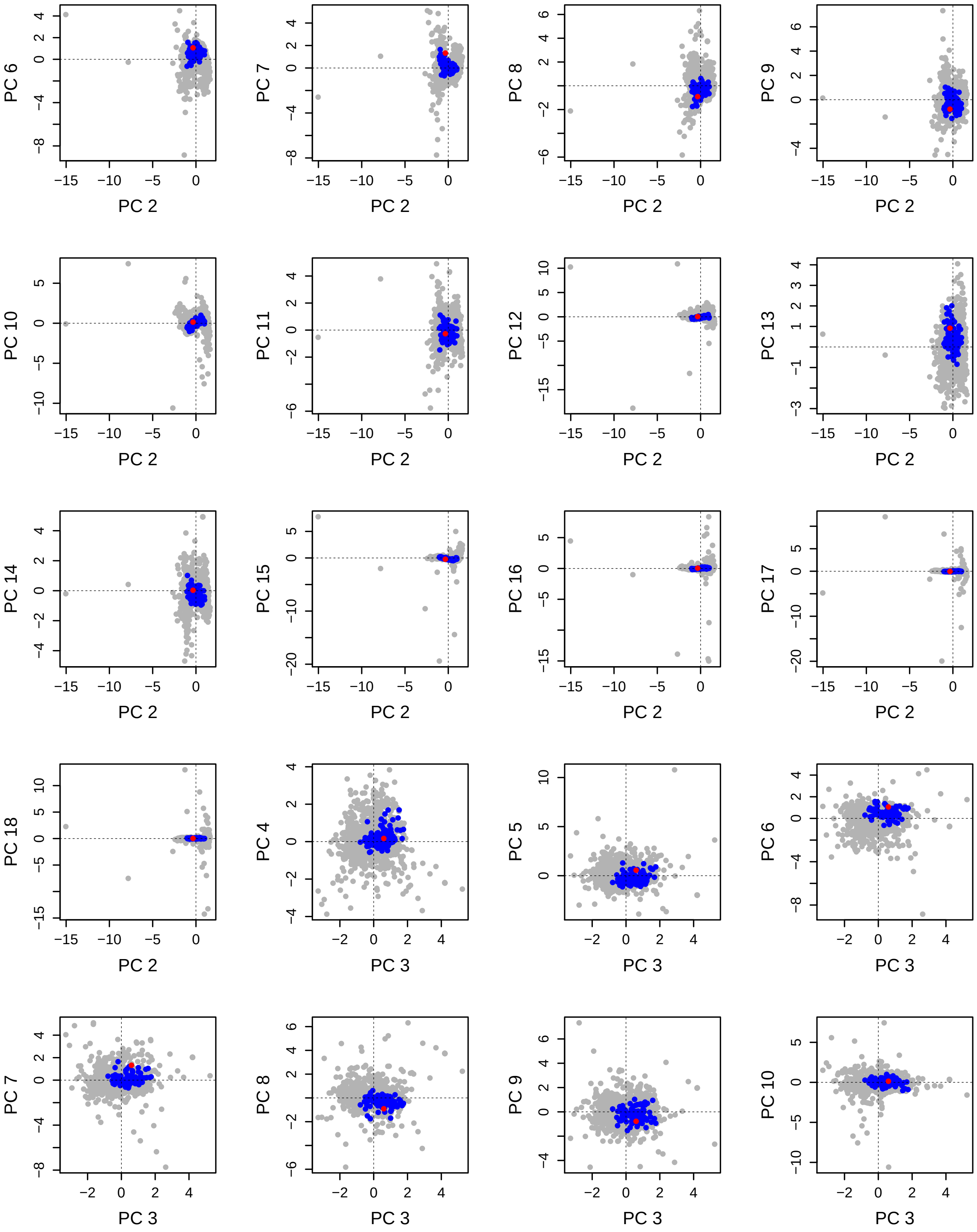


Figure S2 *Continued on next page*

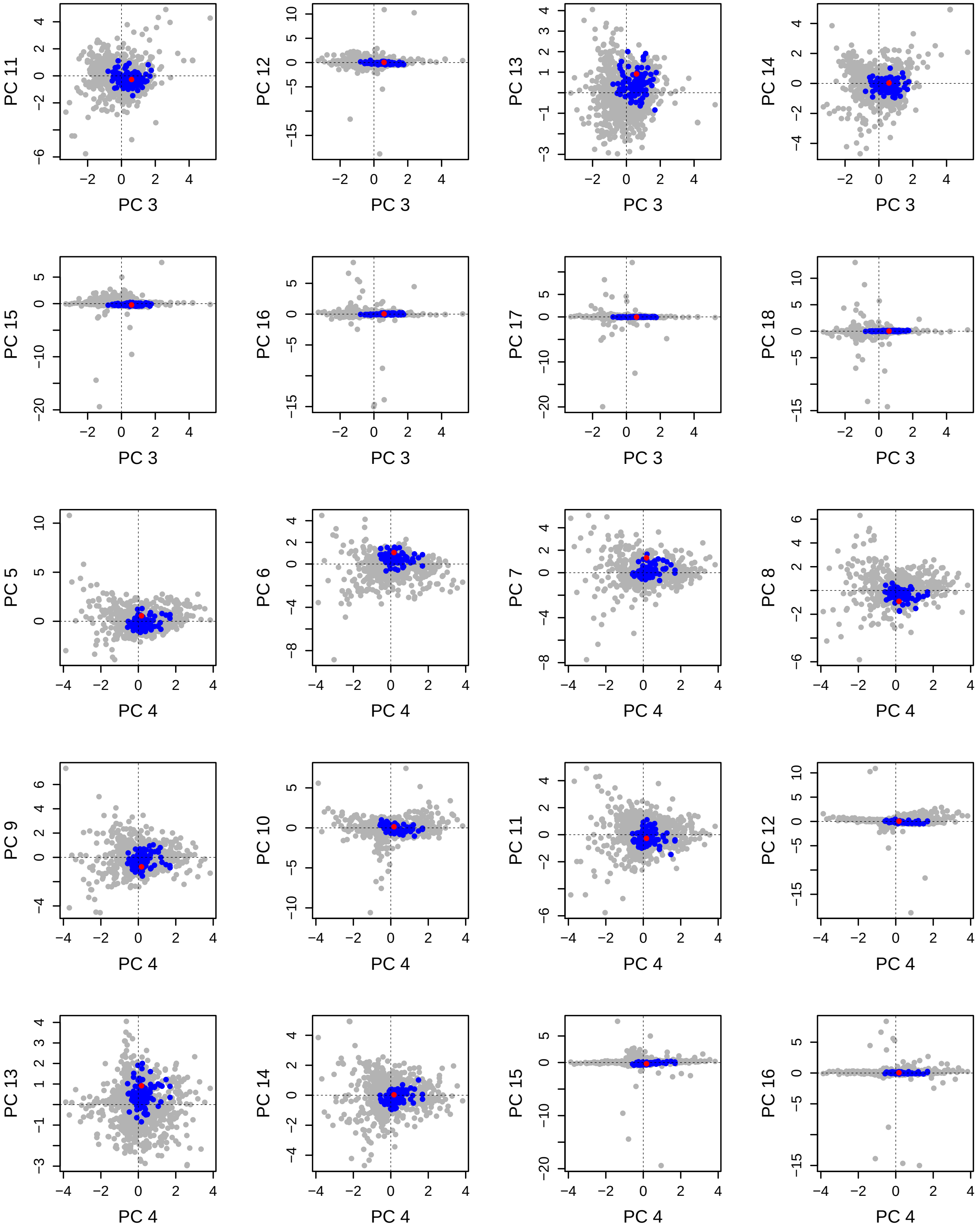


Figure S2 *Continued on next page*

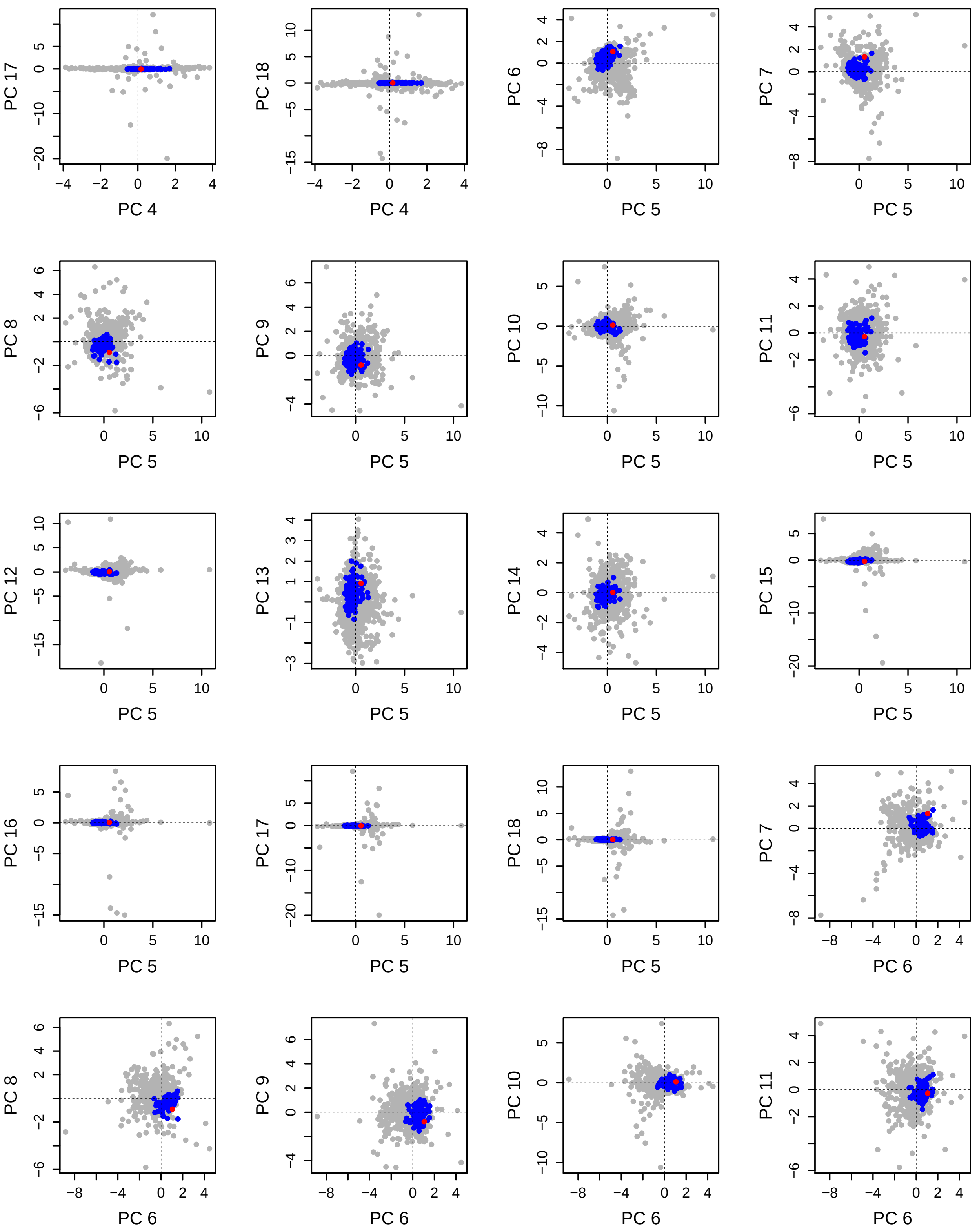


Figure S2 *Continued on next page*

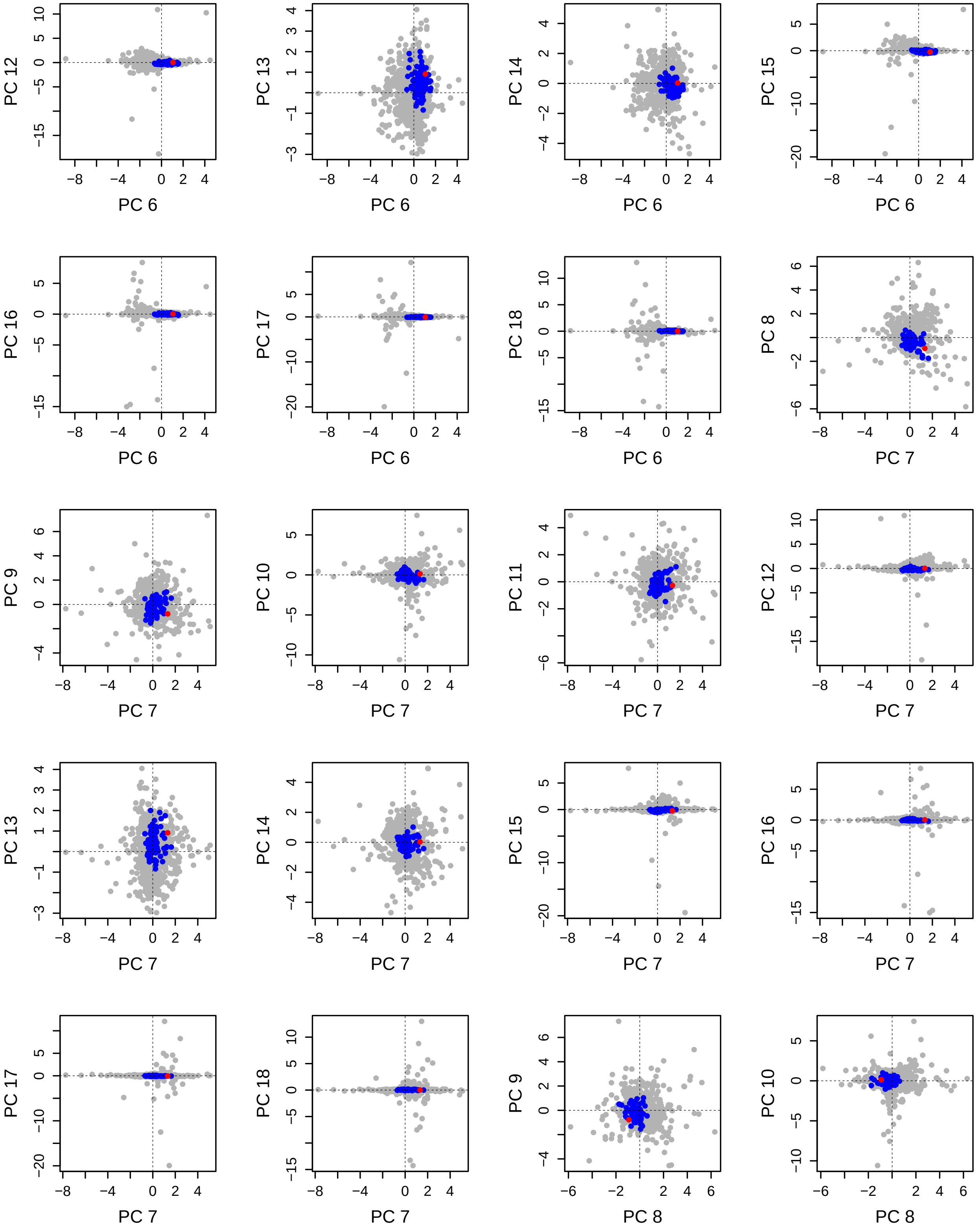


Figure S2 *Continued on next page*

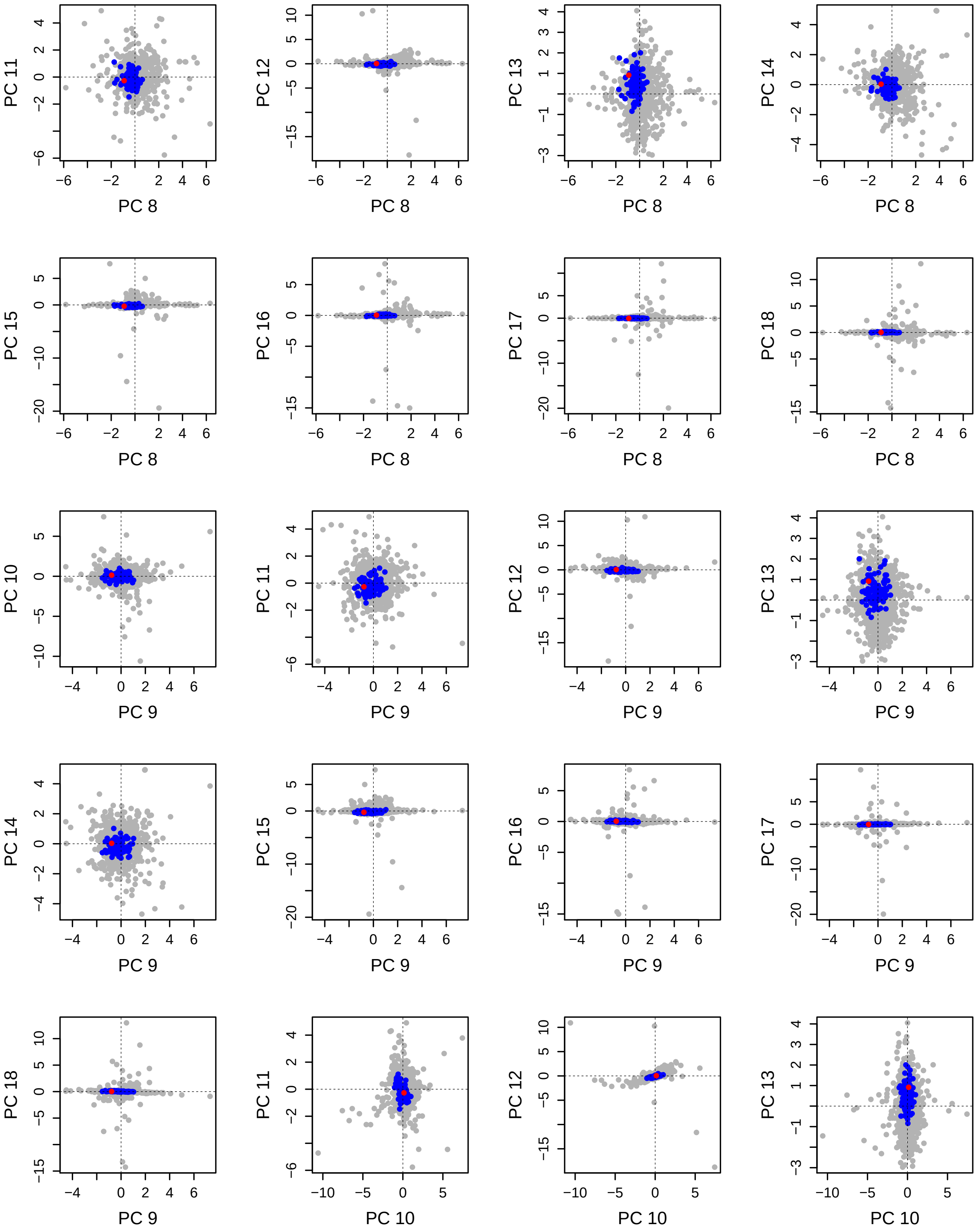


Figure S2 *Continued on next page*

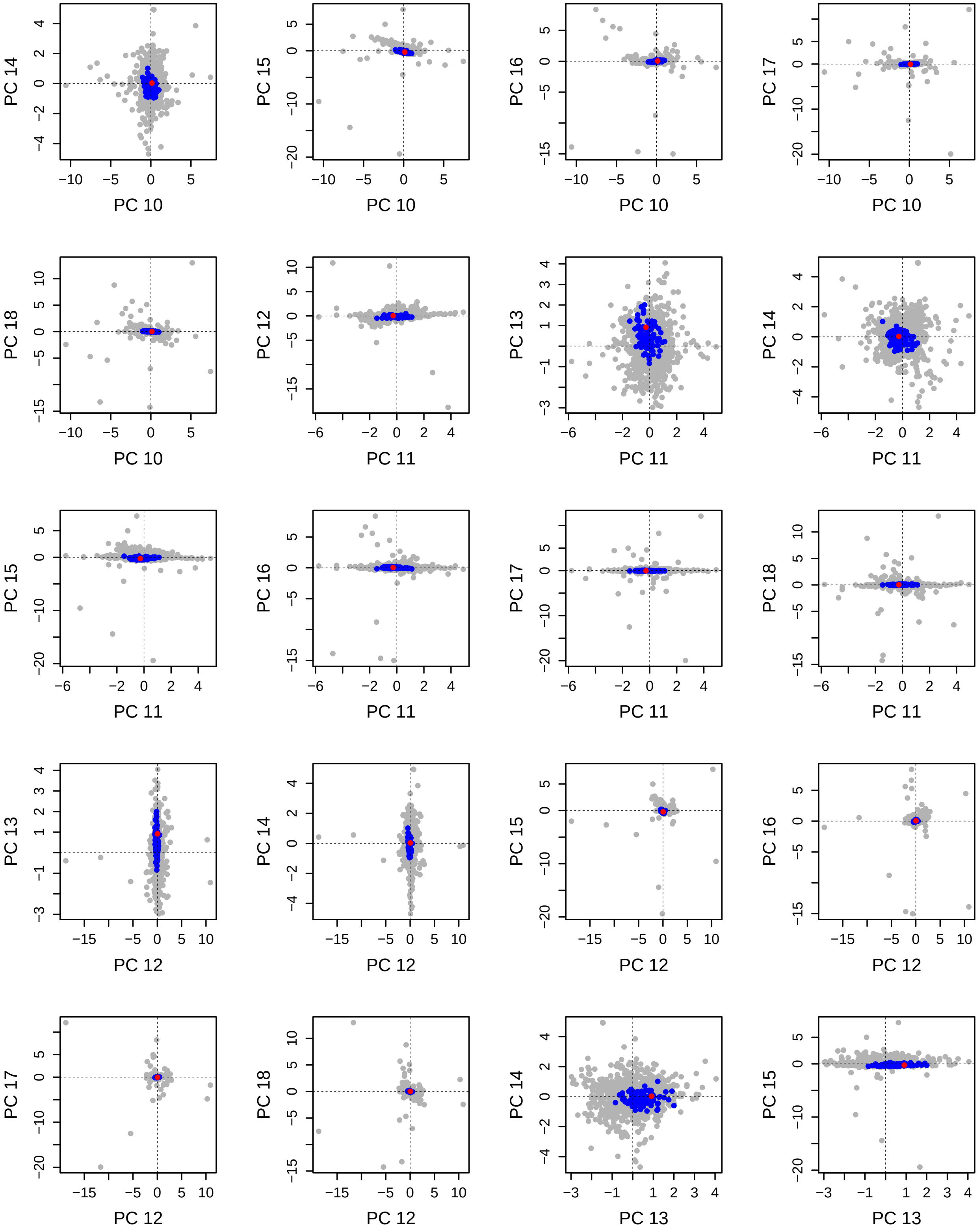


Figure S2 *Continued on next page*

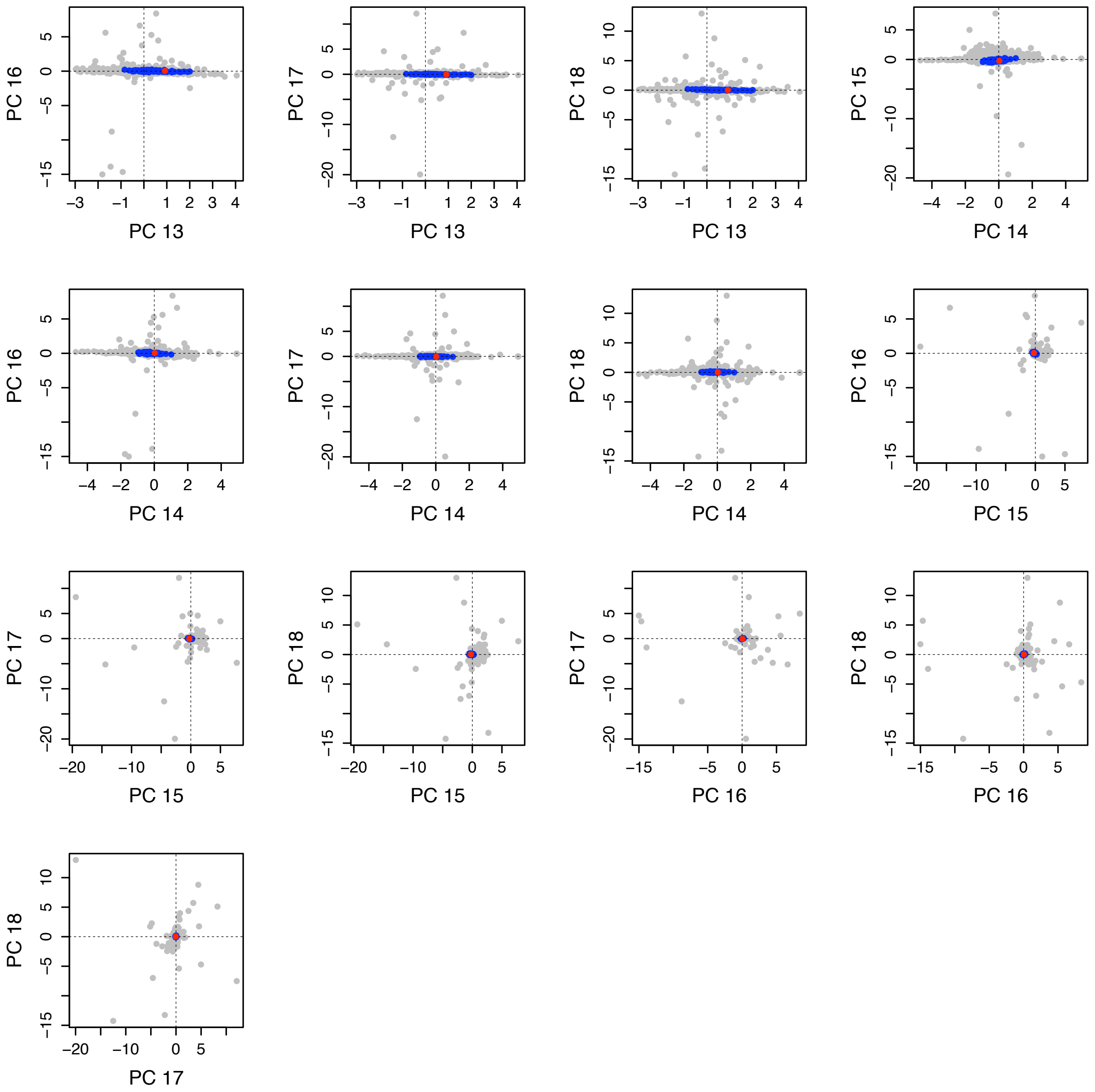


Figure S2 *Continued from previous page.* Pairwise prior predictive distribution of PC-rotated summary statistics. Gray points represent $N = 1000$ simulations with parameter values drawn from the prior. The true value from the ibex data set is shown as a red dot. The fact that it is always embedded in the cloud of gray points means that the model and prior distributions are well specified. The $n' = 100$ points with smallest Euclidean distance from the observation are shown in blue. Those represent simulations used as training data sets for the *local* choice of summary statistics (see main text). In the main study, we used $N = 10^6$ and $n' = 1000$; smaller numbers are used here for illustration of the principle.

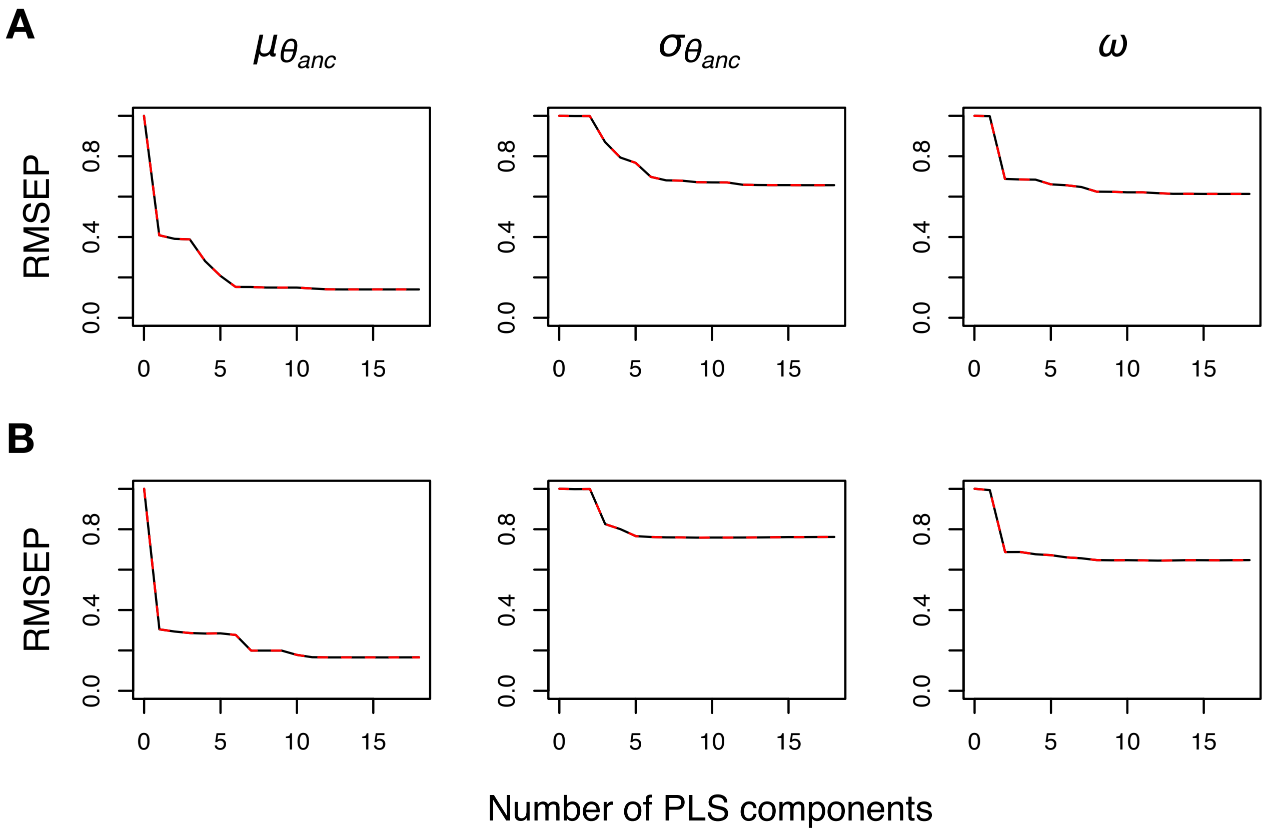


Figure S3 Root mean squared error of prediction (RMSEP) for PLS regression as a function of the number of PLS components used. As suggested by (Wegmann *et al.* 2009), we chose the number of PLS components to be kept as summary statistics based on these plots. The RMSEP was obtained via leave-one-out cross-validation. (A) Global and (B) local choice of summary statistics via PLS (see main text). In (B), the observation from the ibex data set was used as the center. In both cases, we decided to keep the first ten components as summary statistics.

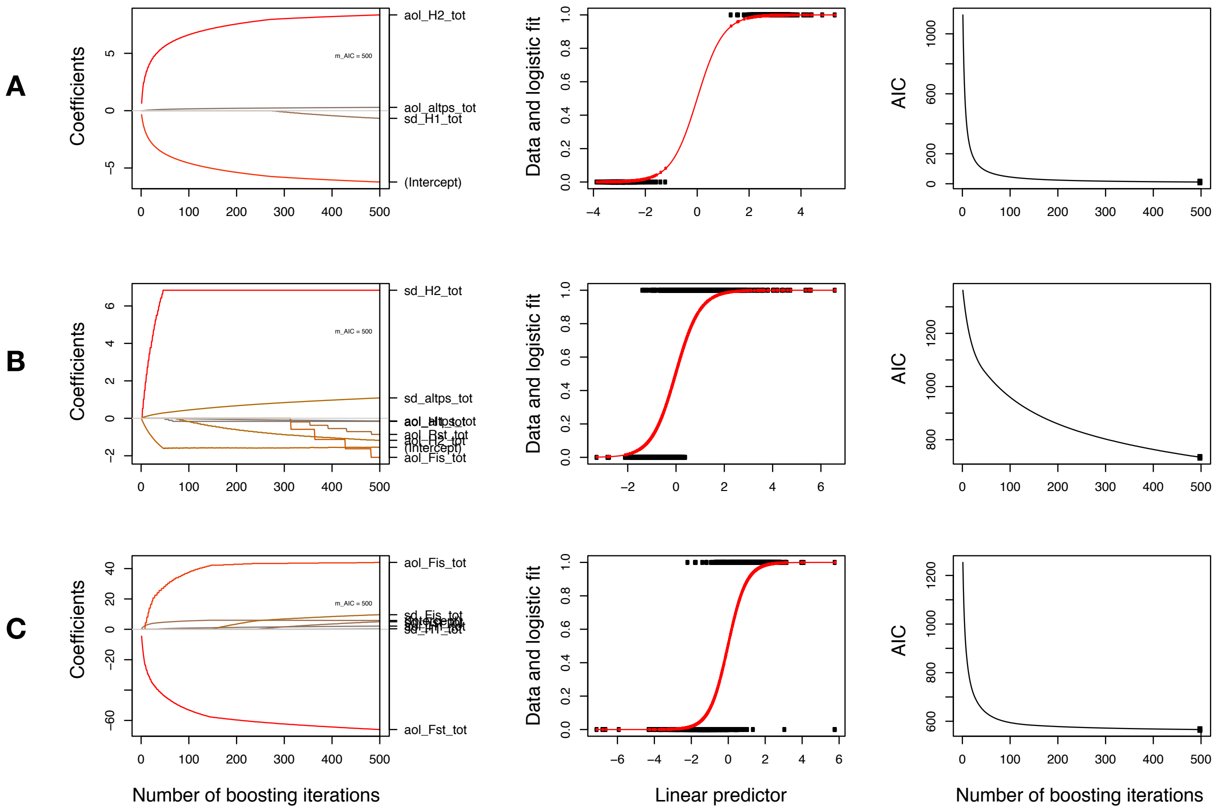


Figure S4 Choice of summary statistics via LogitBoost for the three parameters $\mu_{\theta_{\text{anc}}}$ (A), $\sigma_{\theta_{\text{anc}}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Middle column: Binary parameter class variable (Y, black) and logistic fit to the probability $\Pr[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ (red), as a function of the linear predictor. Right column: Quality of fit in terms of AIC as a function of the number of iterations m. The thick black line marks the m_{stop} chosen. In the cases shown here, no minimum AIC was found for $m < 500$.

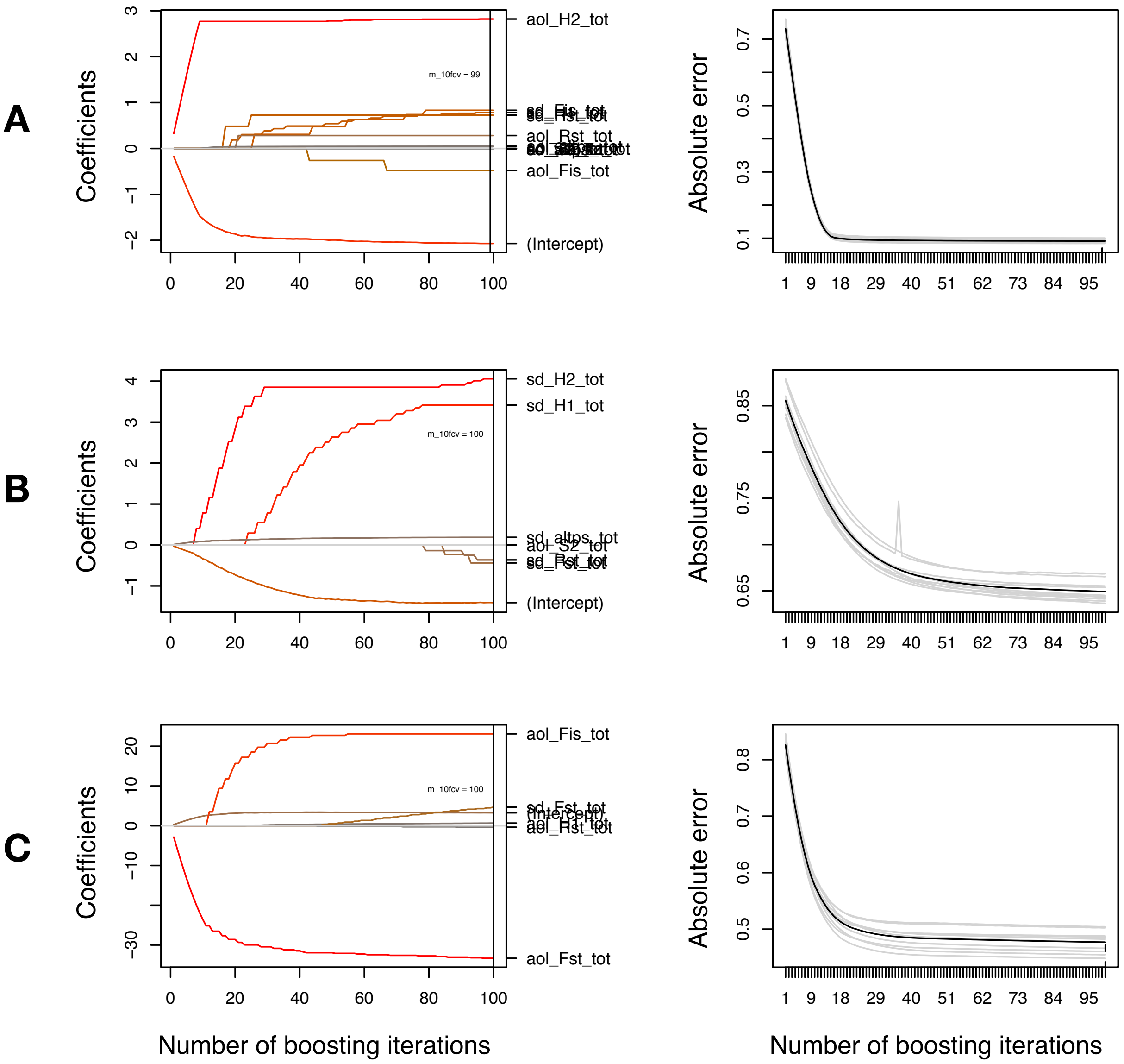


Figure S5 Choice of summary statistics via L_1Boosting for the three parameters $\mu_{\theta_{\rm anc}}$ (A), $\sigma_{\theta_{\rm anc}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Right column: Quality of fit in terms of the bootstrapping error, as a function of the number of iterations m. The dashed vertical line marks the m_{stop} chosen. In the cases shown here, no minimum absolute error was found for $m < 100$.

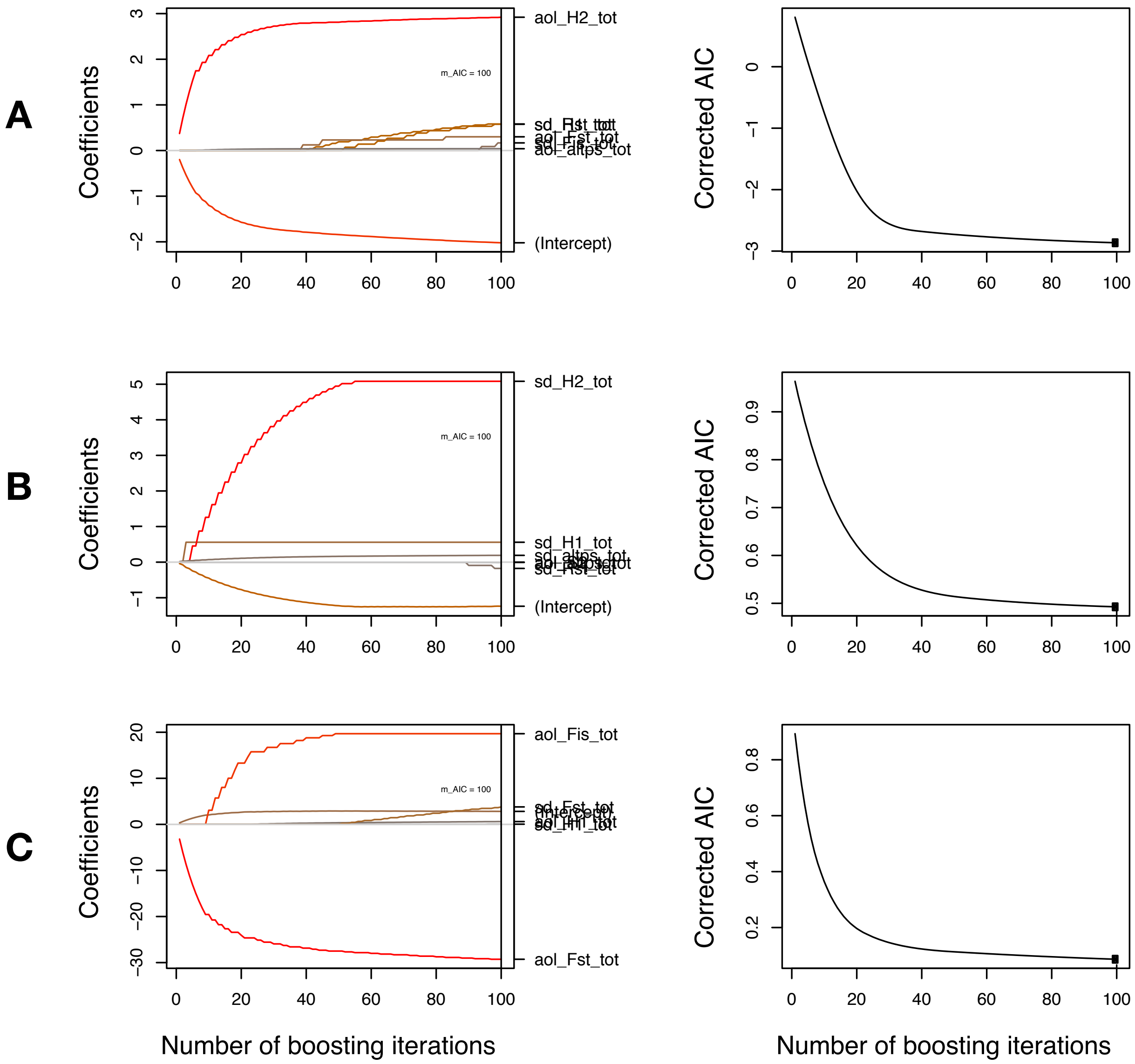


Figure S6 Choice of summary statistics via L_2Boosting for the three parameters $\mu_{\theta_{\text{anc}}}$ (A), $\sigma_{\theta_{\text{anc}}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Right column: Quality of fit in terms of the corrected AIC as a function of the number of iterations m. The thick black line marks the m_{stop} chosen. In the cases shown here, no minimum absolute error was found for $m < 100$.

Figure S7 ***(facing page)*** Posterior distributions inferred for six random test data sets with acceptance rate $\epsilon = 0.01$. Methods are as described in the main text. True values are given by a dashed vertical line, prior distributions in blue (*cf.* Table 1).

ε = 0.01

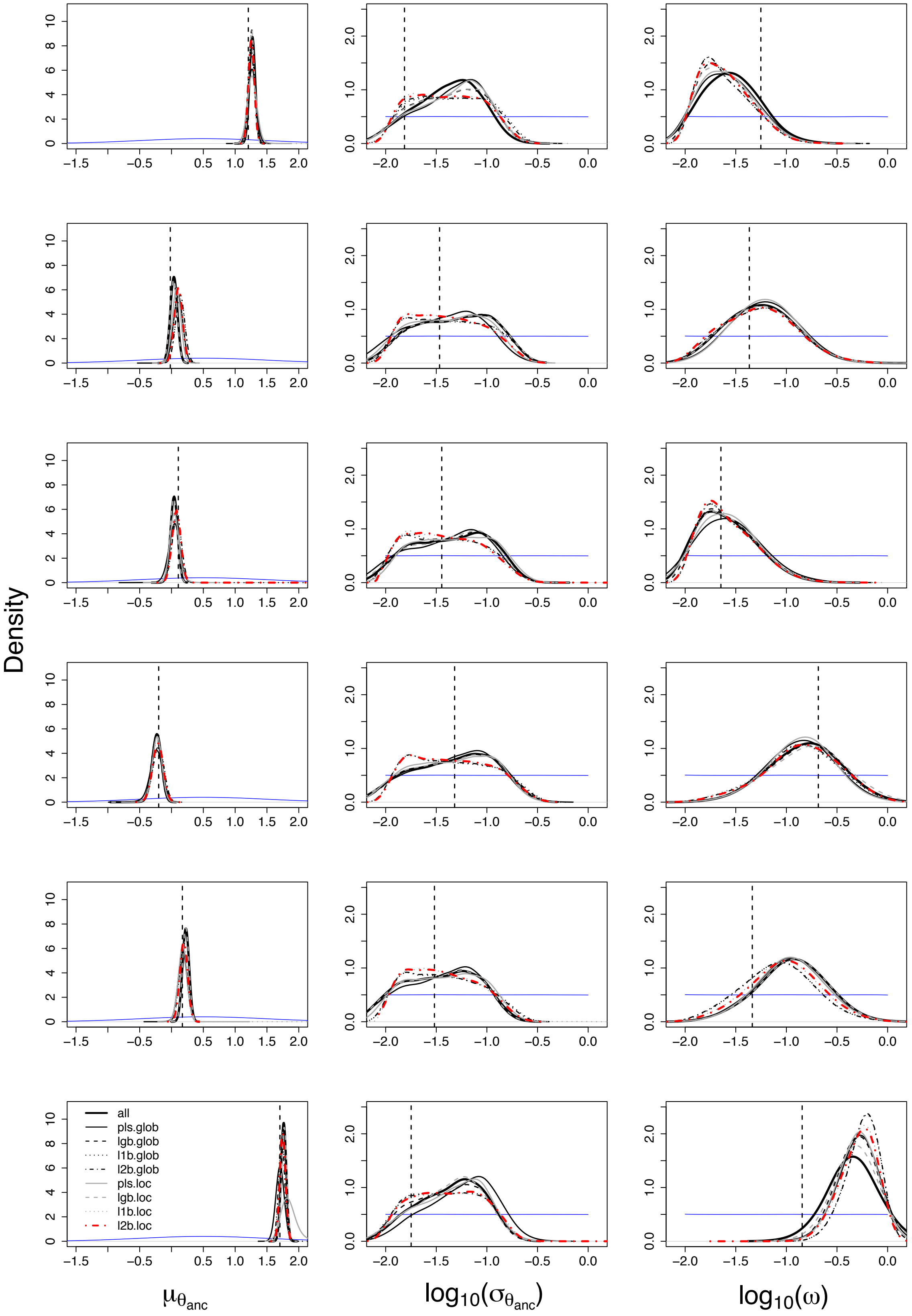

Density
0
2
4
6
8
10
0.0
1.0
2.0
−1.5
−0.5
0.5
1.0
1.5
2.0
−2.0
−1.5
−1.0
−0.5
0.0
all
pls.glob
lgb.glob
l1b.glob
l2b.glob
pls.loc
lgb.loc
l1b.loc
l2b.loc
$\mu_{\theta_{anc}}$
$\log_{10}(\sigma_{\theta_{anc}})$
$\log_{10}(\omega)$

Figure S8 ***(facing page)*** Coverage property of posterior distributions inferred with different choices of summary statistics on a global scale. Histograms show the distribution across 500 independent test estimations of the posterior probabilities of the true parameter values. The distribution is expected to be uniform (Wegmann *et al.* 2009). Left-skewed or right-skewed distributions indicate that the parameter is on average over- or underestimated, respectively. Peaked or U-shaped distributions result from posterior distributions that are too wide or too narrow, respectively. Non-uniform distributions are shaded in gray (p-values from a Kolmogorov-Smirnov test on top are without correction for multiple testing; see main text).

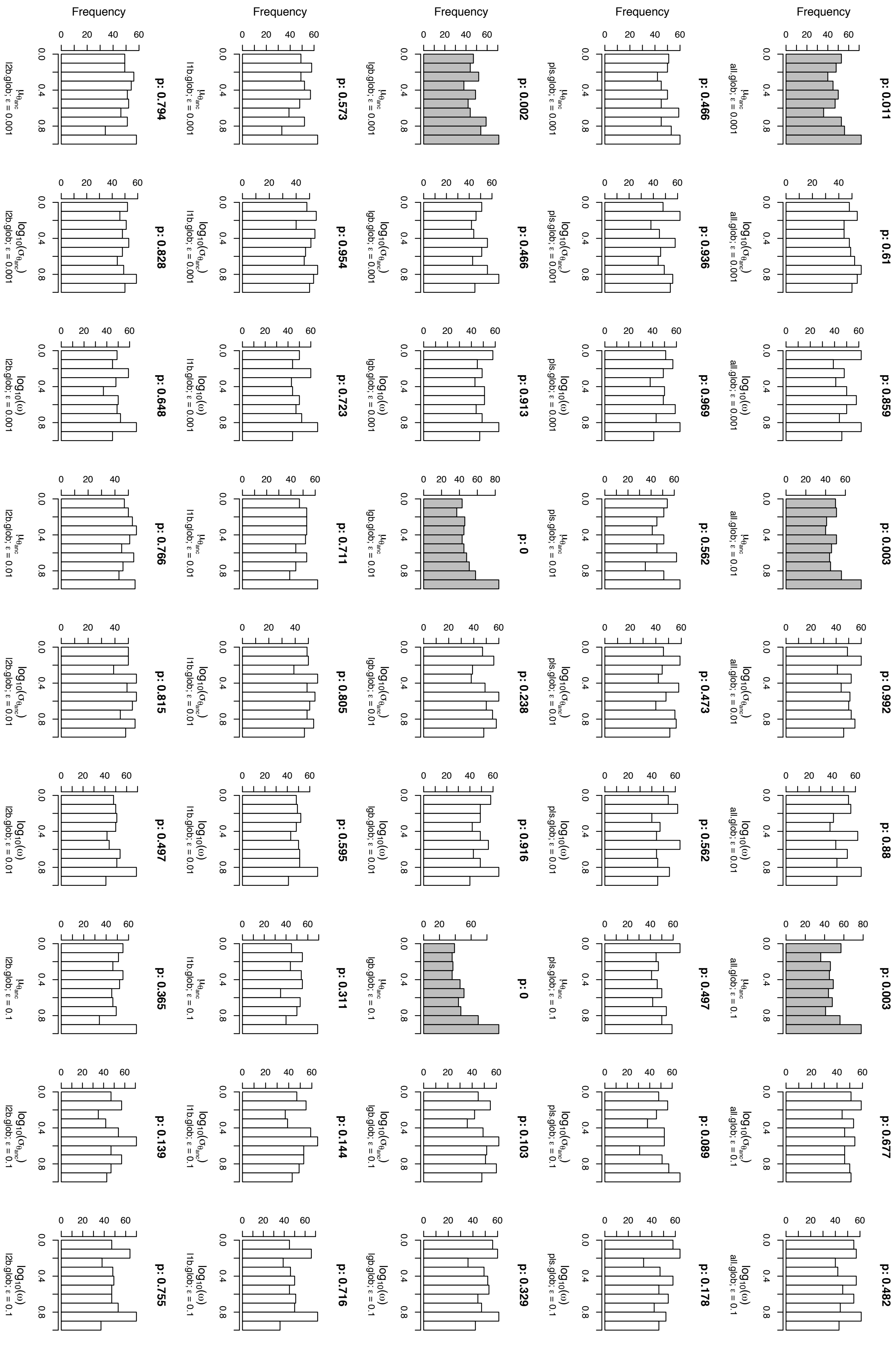

Frequency
p: 0.011
$\mu_{\theta_{anc}}$
all.glob; ε = 0.001
p: 0.61
$\log_{10}(\sigma_{\theta_{anc}})$
all.glob; ε = 0.001
p: 0.859
$\log_{10}(\omega)$
all.glob; ε = 0.001
p: 0.003
$\mu_{\theta_{anc}}$
all.glob; ε = 0.01
p: 0.992
$\log_{10}(\sigma_{\theta_{anc}})$
all.glob; ε = 0.01
p: 0.88
$\log_{10}(\omega)$
all.glob; ε = 0.01
p: 0.003
$\mu_{\theta_{anc}}$
all.glob; ε = 0.1
p: 0.677
$\log_{10}(\sigma_{\theta_{anc}})$
all.glob; ε = 0.1
p: 0.482
$\log_{10}(\omega)$
all.glob; ε = 0.1
Frequency
p: 0.466
$\mu_{\theta_{anc}}$
pls.glob; ε = 0.001
p: 0.936
$\log_{10}(\sigma_{\theta_{anc}})$
pls.glob; ε = 0.001
p: 0.969
$\log_{10}(\omega)$
pls.glob; ε = 0.001
p: 0.562
$\mu_{\theta_{anc}}$
pls.glob; ε = 0.01
p: 0.473
$\log_{10}(\sigma_{\theta_{anc}})$
pls.glob; ε = 0.01
p: 0.562
$\log_{10}(\omega)$
pls.glob; ε = 0.01
p: 0.497
$\mu_{\theta_{anc}}$
pls.glob; ε = 0.1
p: 0.089
$\log_{10}(\sigma_{\theta_{anc}})$
pls.glob; ε = 0.1
p: 0.178
$\log_{10}(\omega)$
pls.glob; ε = 0.1
Frequency
p: 0.002
$\mu_{\theta_{anc}}$
lgb.glob; ε = 0.001
p: 0.466
$\log_{10}(\sigma_{\theta_{anc}})$
lgb.glob; ε = 0.001
p: 0.913
$\log_{10}(\omega)$
lgb.glob; ε = 0.001
p: 0
$\mu_{\theta_{anc}}$
lgb.glob; ε = 0.01
p: 0.238
$\log_{10}(\sigma_{\theta_{anc}})$
lgb.glob; ε = 0.01
p: 0.916
$\log_{10}(\omega)$
lgb.glob; ε = 0.01
p: 0
$\mu_{\theta_{anc}}$
lgb.glob; ε = 0.1
p: 0.103
$\log_{10}(\sigma_{\theta_{anc}})$
lgb.glob; ε = 0.1
p: 0.329
$\log_{10}(\omega)$
lgb.glob; ε = 0.1
Frequency
p: 0.573
$\mu_{\theta_{anc}}$
l1b.glob; ε = 0.001
p: 0.954
$\log_{10}(\sigma_{\theta_{anc}})$
l1b.glob; ε = 0.001
p: 0.723
$\log_{10}(\omega)$
l1b.glob; ε = 0.001
p: 0.711
$\mu_{\theta_{anc}}$
l1b.glob; ε = 0.01
p: 0.805
$\log_{10}(\sigma_{\theta_{anc}})$
l1b.glob; ε = 0.01
p: 0.595
$\log_{10}(\omega)$
l1b.glob; ε = 0.01
p: 0.311
$\mu_{\theta_{anc}}$
l1b.glob; ε = 0.1
p: 0.144
$\log_{10}(\sigma_{\theta_{anc}})$
l1b.glob; ε = 0.1
p: 0.716
$\log_{10}(\omega)$
l1b.glob; ε = 0.1
Frequency
p: 0.794
$\mu_{\theta_{anc}}$
l2b.glob; ε = 0.001
p: 0.828
$\log_{10}(\sigma_{\theta_{anc}})$
l2b.glob; ε = 0.001
p: 0.648
$\log_{10}(\omega)$
l2b.glob; ε = 0.001
p: 0.766
$\mu_{\theta_{anc}}$
l2b.glob; ε = 0.01
p: 0.815
$\log_{10}(\sigma_{\theta_{anc}})$
l2b.glob; ε = 0.01
p: 0.497
$\log_{10}(\omega)$
l2b.glob; ε = 0.01
p: 0.365
$\mu_{\theta_{anc}}$
l2b.glob; ε = 0.1
p: 0.139
$\log_{10}(\sigma_{\theta_{anc}})$
l2b.glob; ε = 0.1
p: 0.755
$\log_{10}(\omega)$
l2b.glob; ε = 0.1
0.0
0.4
0.8
0
20
40
60
80

Figure S9 ***(facing page)*** Coverage property of posterior distributions inferred with different choices of summary statistics on a local scale. Non-uniform distributions of posterior probabilities are shaded in gray (p-values from a Kolmogorov-Smirnov test on top). Note that the first row here corresponds to the first row in Figure S8. Further details as in Figure S8.

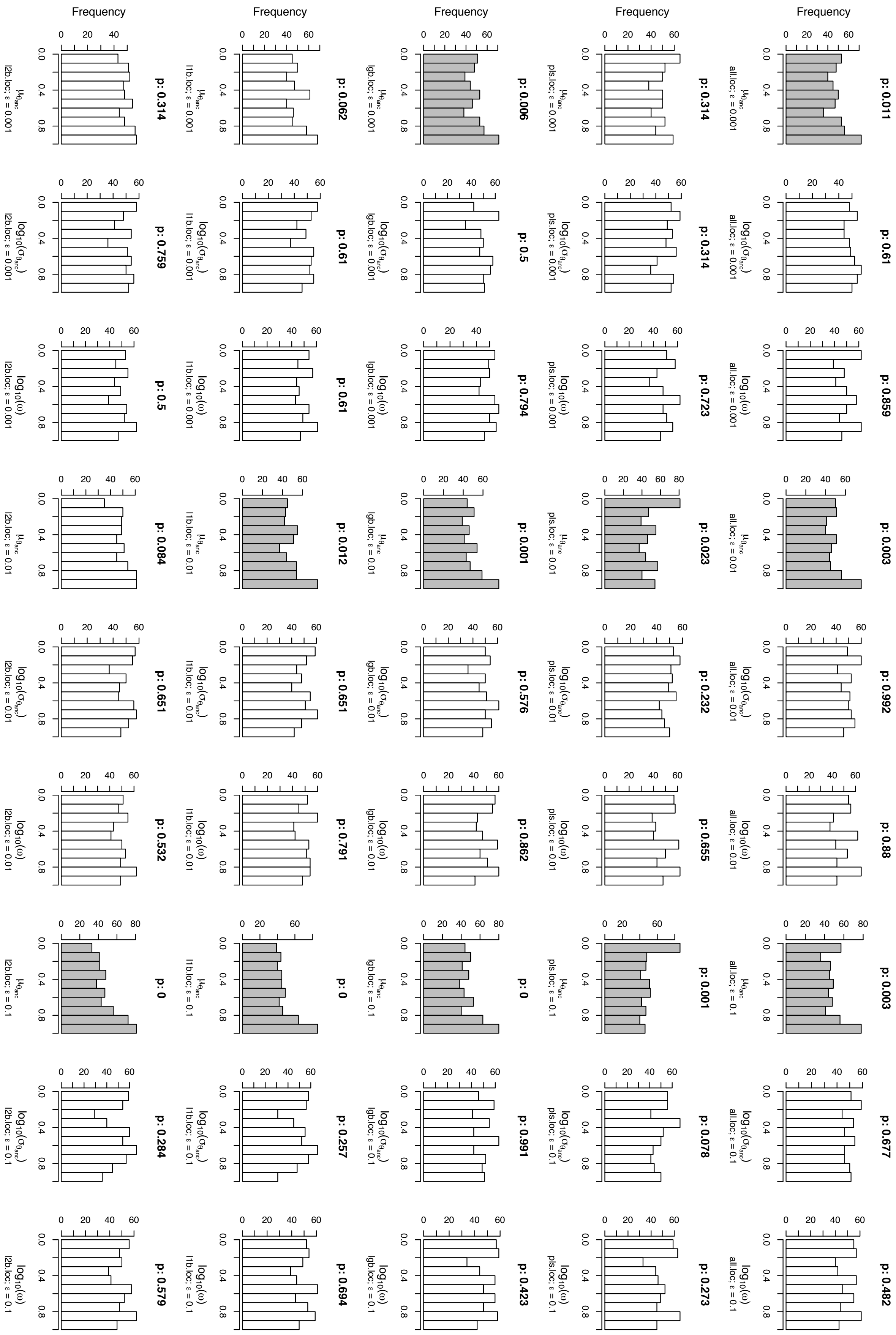
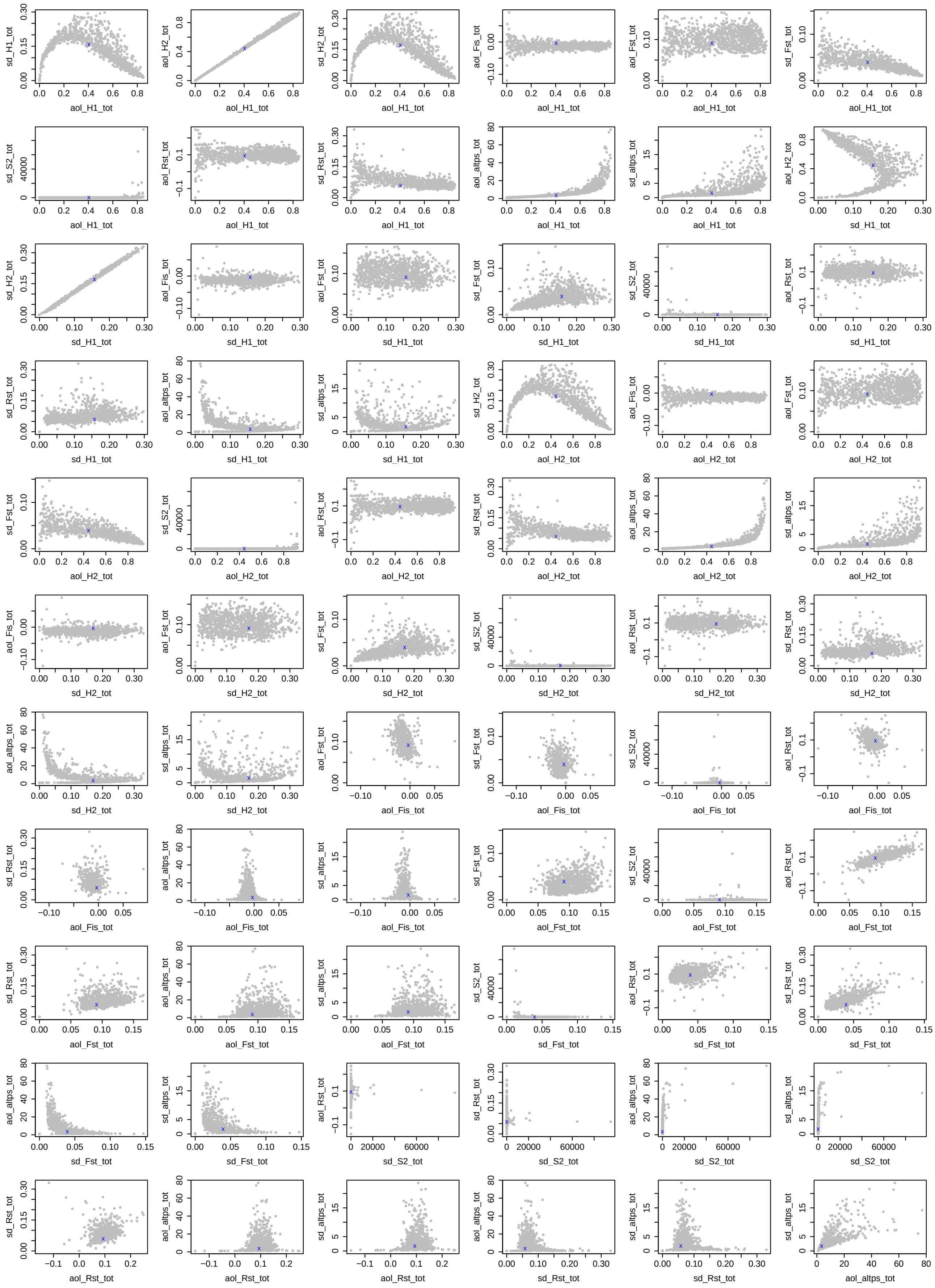

Frequency
0 20 40 60 80
0.0 0.4 0.8
p: 0.011
$\mu_{\theta_{anc}}$ all.loc; ε = 0.001
p: 0.61
$\log_{10}(\sigma_{\theta_{anc}})$ all.loc; ε = 0.001
p: 0.859
$\log_{10}(\omega)$ all.loc; ε = 0.001
p: 0.003
$\mu_{\theta_{anc}}$ all.loc; ε = 0.01
p: 0.992
$\log_{10}(\sigma_{\theta_{anc}})$ all.loc; ε = 0.01
p: 0.88
$\log_{10}(\omega)$ all.loc; ε = 0.01
p: 0.003
$\mu_{\theta_{anc}}$ all.loc; ε = 0.1
p: 0.677
$\log_{10}(\sigma_{\theta_{anc}})$ all.loc; ε = 0.1
p: 0.482
$\log_{10}(\omega)$ all.loc; ε = 0.1
p: 0.314
$\mu_{\theta_{anc}}$ pls.loc; ε = 0.001
p: 0.314
$\log_{10}(\sigma_{\theta_{anc}})$ pls.loc; ε = 0.001
p: 0.723
$\log_{10}(\omega)$ pls.loc; ε = 0.001
p: 0.023
$\mu_{\theta_{anc}}$ pls.loc; ε = 0.01
p: 0.232
$\log_{10}(\sigma_{\theta_{anc}})$ pls.loc; ε = 0.01
p: 0.655
$\log_{10}(\omega)$ pls.loc; ε = 0.01
p: 0.001
$\mu_{\theta_{anc}}$ pls.loc; ε = 0.1
p: 0.078
$\log_{10}(\sigma_{\theta_{anc}})$ pls.loc; ε = 0.1
p: 0.273
$\log_{10}(\omega)$ pls.loc; ε = 0.1
p: 0.006
$\mu_{\theta_{anc}}$ igb.loc; ε = 0.001
p: 0.5
$\log_{10}(\sigma_{\theta_{anc}})$ igb.loc; ε = 0.001
p: 0.794
$\log_{10}(\omega)$ igb.loc; ε = 0.001
p: 0.001
$\mu_{\theta_{anc}}$ igb.loc; ε = 0.01
p: 0.576
$\log_{10}(\sigma_{\theta_{anc}})$ igb.loc; ε = 0.01
p: 0.862
$\log_{10}(\omega)$ igb.loc; ε = 0.01
p: 0
$\mu_{\theta_{anc}}$ igb.loc; ε = 0.1
p: 0.991
$\log_{10}(\sigma_{\theta_{anc}})$ igb.loc; ε = 0.1
p: 0.423
$\log_{10}(\omega)$ igb.loc; ε = 0.1
p: 0.062
$\mu_{\theta_{anc}}$ l1b.loc; ε = 0.001
p: 0.61
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.loc; ε = 0.001
p: 0.61
$\log_{10}(\omega)$ l1b.loc; ε = 0.001
p: 0.012
$\mu_{\theta_{anc}}$ l1b.loc; ε = 0.01
p: 0.651
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.loc; ε = 0.01
p: 0.791
$\log_{10}(\omega)$ l1b.loc; ε = 0.01
p: 0
$\mu_{\theta_{anc}}$ l1b.loc; ε = 0.1
p: 0.257
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.loc; ε = 0.1
p: 0.694
$\log_{10}(\omega)$ l1b.loc; ε = 0.1
p: 0.314
$\mu_{\theta_{anc}}$ l2b.loc; ε = 0.001
p: 0.759
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.loc; ε = 0.001
p: 0.5
$\log_{10}(\omega)$ l2b.loc; ε = 0.001
p: 0.084
$\mu_{\theta_{anc}}$ l2b.loc; ε = 0.01
p: 0.651
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.loc; ε = 0.01
p: 0.532
$\log_{10}(\omega)$ l2b.loc; ε = 0.01
p: 0
$\mu_{\theta_{anc}}$ l2b.loc; ε = 0.1
p: 0.284
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.loc; ε = 0.1
p: 0.579
$\log_{10}(\omega)$ l2b.loc; ε = 0.1

Figure S10 ***(facing page)*** Pairwise prior predictive distribution of summary statistics on original scale. Only summary statistics chosen with the $\mathrm{lgb.glob}$ method are shown. Gray points represent $N = 1000$ simulations with parameter values drawn from the prior. The true value from the ibex data set is shown as a blue cross; *aol*, average over loci; *sd*, standard deviation over loci.


sd_H1_tot
aol_H1_tot
aol_H2_tot
aol_H1_tot
sd_H2_tot
aol_H1_tot
aol_Fis_tot
aol_H1_tot
aol_Fst_tot
aol_H1_tot
sd_Fst_tot
aol_H1_tot
sd_S2_tot
aol_H1_tot
aol_Rst_tot
aol_H1_tot
sd_Rst_tot
aol_H1_tot
aol_altps_tot
aol_H1_tot
sd_altps_tot
aol_H1_tot
aol_H2_tot
sd_H1_tot
sd_H2_tot
sd_H1_tot
aol_Fis_tot
sd_H1_tot
aol_Fst_tot
sd_H1_tot
sd_Fst_tot
sd_H1_tot
sd_S2_tot
sd_H1_tot
aol_Rst_tot
sd_H1_tot
sd_Rst_tot
sd_H1_tot
aol_altps_tot
sd_H1_tot
sd_altps_tot
sd_H1_tot
sd_H2_tot
aol_H2_tot
aol_Fis_tot
aol_H2_tot
aol_Fst_tot
aol_H2_tot
sd_Fst_tot
aol_H2_tot
sd_S2_tot
aol_H2_tot
aol_Rst_tot
aol_H2_tot
sd_Rst_tot
aol_H2_tot
aol_altps_tot
aol_H2_tot
sd_altps_tot
aol_H2_tot
aol_Fis_tot
sd_H2_tot
aol_Fst_tot
sd_H2_tot
sd_Fst_tot
sd_H2_tot
sd_S2_tot
sd_H2_tot
aol_Rst_tot
sd_H2_tot
sd_Rst_tot
sd_H2_tot
aol_altps_tot
sd_H2_tot
sd_altps_tot
sd_H2_tot
aol_Fst_tot
aol_Fis_tot
sd_Fst_tot
aol_Fis_tot
sd_S2_tot
aol_Fis_tot
aol_Rst_tot
aol_Fis_tot
sd_Rst_tot
aol_Fis_tot
aol_altps_tot
aol_Fis_tot
sd_altps_tot
aol_Fis_tot
sd_Fst_tot
aol_Fst_tot
sd_S2_tot
aol_Fst_tot
aol_Rst_tot
aol_Fst_tot
sd_Rst_tot
aol_Fst_tot
aol_altps_tot
aol_Fst_tot
sd_altps_tot
aol_Fst_tot
sd_S2_tot
sd_Fst_tot
aol_Rst_tot
sd_Fst_tot
sd_Rst_tot
sd_Fst_tot
aol_altps_tot
sd_Fst_tot
sd_altps_tot
sd_Fst_tot
aol_Rst_tot
sd_S2_tot
sd_Rst_tot
sd_S2_tot
aol_altps_tot
sd_S2_tot
sd_altps_tot
sd_S2_tot
sd_Rst_tot
aol_Rst_tot
aol_altps_tot
aol_Rst_tot
sd_altps_tot
aol_Rst_tot
aol_altps_tot
sd_Rst_tot
sd_altps_tot
sd_Rst_tot
sd_altps_tot
aol_altps_tot

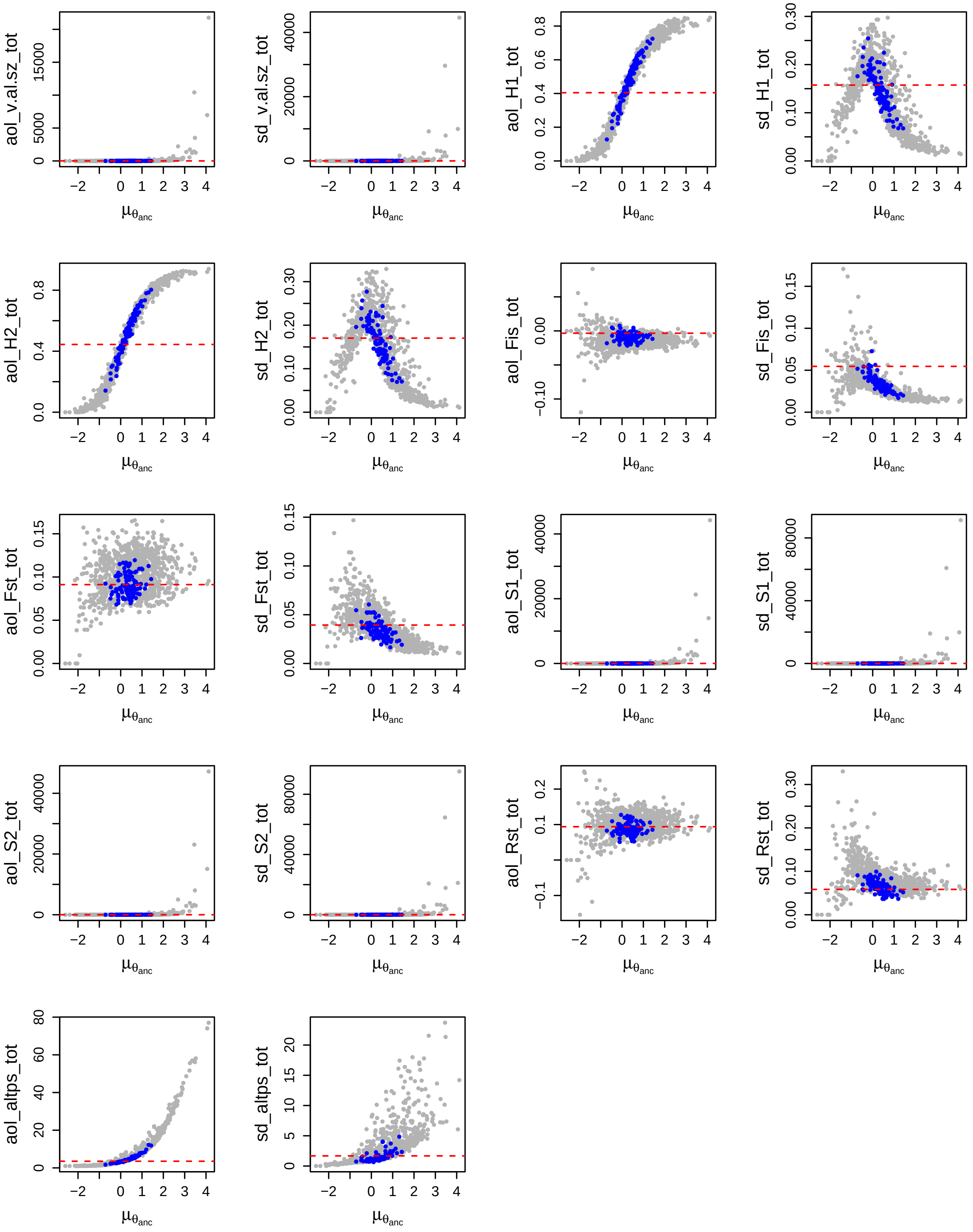


Figure S11 Relation between $\mu_{\theta_{\text{anc}}}$ and the candidate summary statistics. Summary statistics are on the y-axis; *aol*, average over loci; *sd*, standard deviation over loci. Gray points represent $n = 10,000$ simulations, the red dashed line corresponds to the observation for Alpine ibex. Blue points give the $n' = 1000$ simulations closest to the observation, where ‘closeness’ was defined as described in the main text (*cf.* Figure S2).

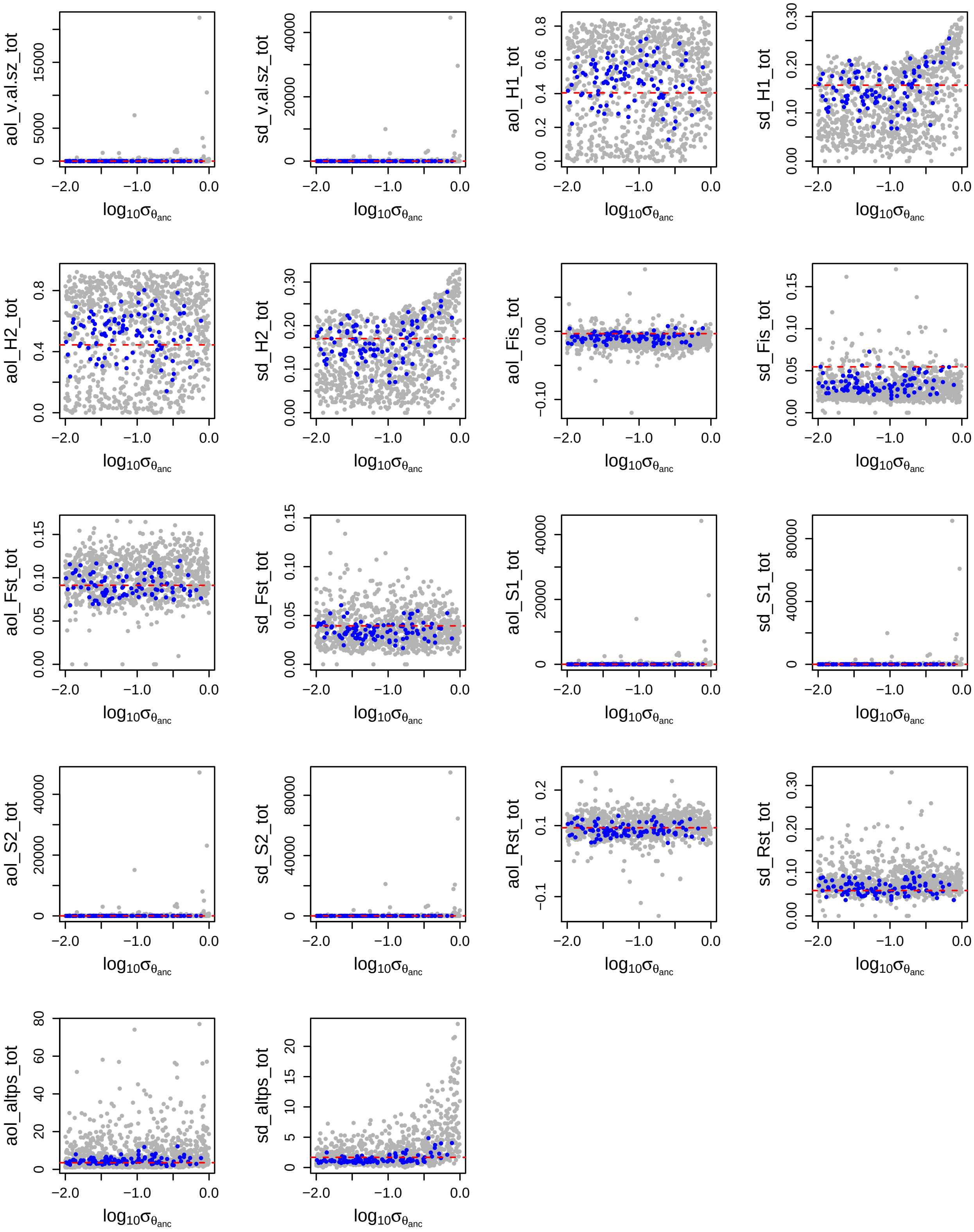


Figure S12 Relation between $\log_{10} \sigma_{\theta_{\text{anc}}}$ and the candidate summary statistics. Details as in Figure S11.

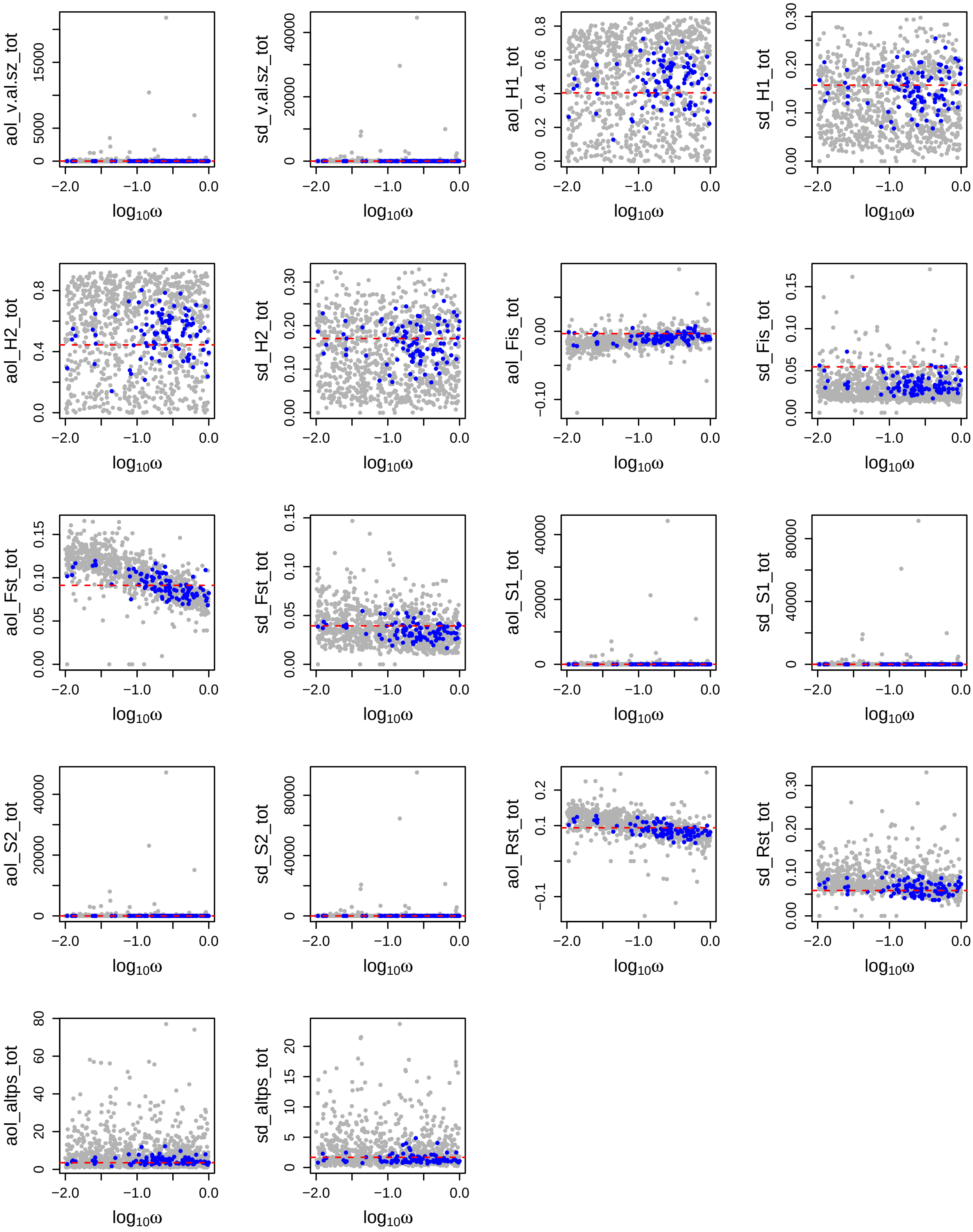


Figure S13 Relation between $\log_{10} \omega$ and the candidate summary statistics. Details as in Figure S11.

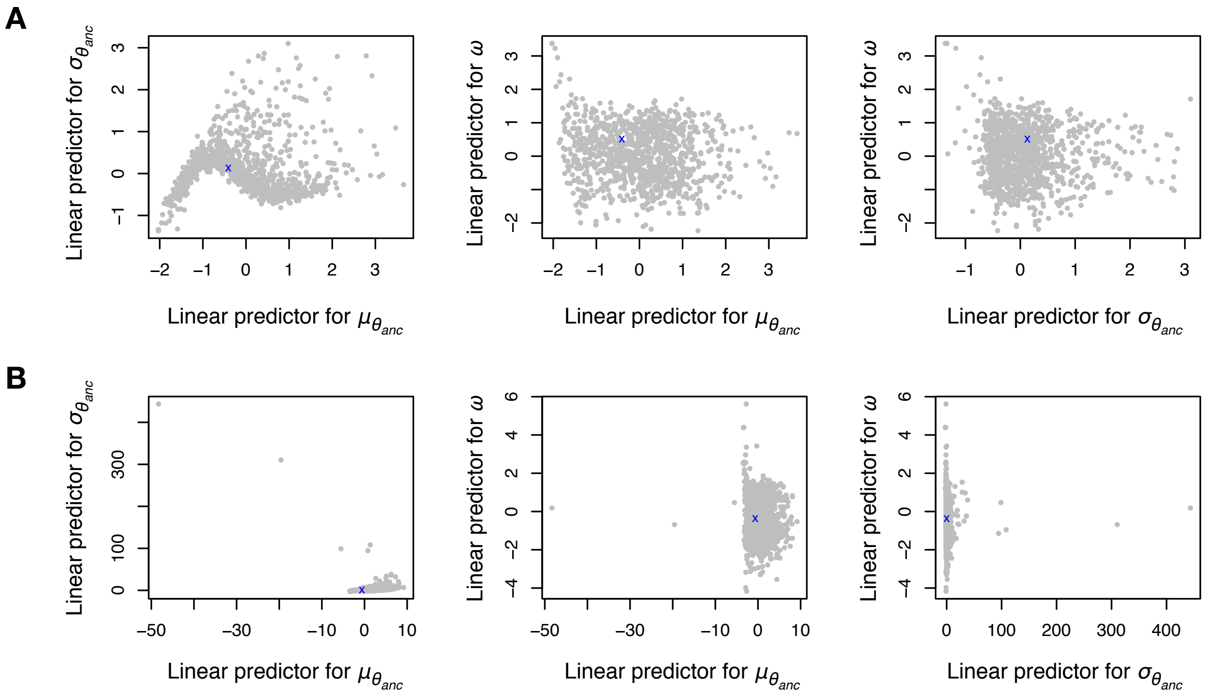


Figure S14 Effect of local choice on scale of summary statistics. Summary statistics were chosen with L_2Boosting as explained in the main text. For each parameter, one linear combination of the original statistics is used as the new summary statistic. These linear combinations are plotted against each other. (A) Global choice of summary statistics. (B) Local choice of summary statistics. Gray points represent $N = 1000$ simulations and the blue cross marks the value observed for Alpine ibex. The local choice of statistics leads to a rescaling compared to the global choice.

SI Supporting Information: Additional Methods

SI.1 Demography and life cycle in simulations

In the following, we give additional details of the demographic model and the ibex-specific settings used in the simulations. All of this is implemented in the program SPoCS (Simulate Populations under Complex Scenarios) written in Java and available on the website http://pub.ist.ac.at/~saeschbacher/phd_e-sources/.

SI.1.1 Life cycle

Alpine ibex is a long-lived, middle-sized ungulate species (Toïgo *et al.* 2002; 2007). We divide the life cycle into years and a year into discrete events, some of which are further described below. We set the maximum age of females and males to 22 and 17 years, respectively (Nievergelt 1966, Toïgo *et al.* 2007). Females and males reach sexual maturity at an age of 3 years (Nievergelt 1966, Stuwe and Grodinsky 1987, Toïgo *et al.* 2002), and the expected age of first reproduction for females and males is 4 and 9 years, respectively (Loison *et al.* 2002, Toïgo *et al.* 2002). In our simulations, females and males stop reproducing when older than 20 and 15 years, respectively.

SI.1.2 Founder/admixture events

A new deme is established by founder individuals taken from previously existing demes. The minimum and maximum age of a founder is 1 and 7 years, respectively, independently of sex. Existing demes may receive further individuals from other demes at later points in time (as specified in Supporting File S3 *transfers*). The range of ages allowed for these admixing individuals is the same as for founders. Founder/admixture events take place at the beginning of the year, before the regulating deaths (see below).

SI.1.3 Reproduction

Females reproduce according to a baseline fertility parameter f. It gives the probability that, for a given year, a particular female will reproduce. If the female reproduces, she mates with a male randomly chosen from the set of males with access to matings in that year (see below). Given a particular female reproduces, it may have one or two offspring. This is controlled by the twin rate parameter $z := \Pr[\text{twins} \mid \text{female reproduces}]$. We set $f = 0.4$ (Nievergelt 1966, Stuwe and Grodinsky 1987) and $z = 0.08$ (Toïgo *et al.* 2002).

Males can get access to matings if they reached the expected age of first reproduction (9 years) and are then counted as potentially reproducing. If, in a deme, no males older than 9 years are available, all males older than the age of sexual maturity (3 years) are considered potentially reproducing. The proportion of these potentially reproducing males that actually get access to matings is defined as ω (see main text). It is one of the parameters to be estimated in this study.

SI.1.4 Deme size control

If the number of offspring required to reach the deme size of the next year cannot be produced by the female baseline fertility f (see above), additional females are allowed to reproduce: Rather than allowing only females to reproduce who reached the expected age of first reproduction (4 years), all females who reached the age of sexual maturity (3 years) may reproduce in this case. If, on the other hand, baseline reproduction results in more individuals than needed to reach the census size of the next year, surplus individuals are removed. These regulating deaths are irrespective of age and sex, and additional to the natural deaths of senescence. In any case, we limit the proportion by which the reproductive need may be overshot per year to 0.2.

SI.1.5 Migration

We simulate migration after the regulating deaths, but before reproduction. Females and males must have reached the age of 3 years before they emigrate (they are then 'potential emigrants'). For a given source deme, the total of individuals to be sent to all connected demes (see main text) are put into an emigrant pool. Emigrants are then randomly distributed to the receiver demes in proportions corresponding to the emigration rates.

SI.2 Explicit forms of minimum expected loss and negative gradient in boosting

The FGD algorithm given in the Appendix of the main text is generic. It is instructive to study the explicit form of expressions in step 1 and 2 of this algorithm for the specific loss functions used here. To this purpose, we follow Friedman *et al.* (2000), Friedman (2001) and Bühlmann and Hothorn (2007).

SI.2.1 Population minimizer of expected loss

We first give explicit forms of the population minimizer (6) for the three loss functions in equations (9), (10) and (12). These are obtained by minimizing the expectation of the joint distribution of $\mathbf{X}$ and Y, $\mathbb{E}_{\mathbf{X},Y}[L(Y,F)]$, where $L(\cdot,\cdot)$ is the generic loss function and $F = F(\mathbf{X})$. In our context, it is enough to take the expectation conditional on $\mathbf{X} = \mathbf{x}$, $\mathbb{E}_Y[L(Y,F) \mid \mathbf{x}]$.

For the L_1-loss in (9), $F^*(\cdot)$ from (6) is obtained as the $F(\cdot)$ that minimizes $\mathbb{E}_Y[|Y - F| \mid \mathbf{x}]$. By the definition of the median, the population minimizer is (Friedman 2001, Bühlmann and Hothorn 2007)

$$F^*(\mathbf{x}) = \mathrm{median}(Y \mid \mathbf{x}). \tag{23}$$

For the L_2-loss in (10), the expected loss is $\mathbb{E}_Y[(Y - F)^2/2 \mid \mathbf{x}]$, and $F^*(\cdot)$ is obtained by setting the derivative with respect to F to zero:

$$\begin{aligned}\frac{\partial}{\partial F}\,\mathbb{E}_Y\left[\frac{1}{2}\,(Y-F)^2\,\middle|\,\mathbf{x}\right] &= \frac{1}{2}\,\frac{\partial\mathbb{E}_Y[Y^2\,|\,\mathbf{x}]}{\partial F} - \frac{\partial\mathbb{E}_Y[Y\,F\,|\,\mathbf{x}]}{\partial F} + \frac{1}{2}\,\frac{\partial\mathbb{E}_Y[F^2\,|\,\mathbf{x}]}{\partial F} \\ &= 0 - \mathbb{E}_Y[Y\,|\,\mathbf{x}] + F(\mathbf{x}) = 0,\end{aligned} \tag{24}$$

from which the familiar result

$$F^*(\mathbf{x}) = \mathbb{E}_Y[Y\,|\,\mathbf{x}] \tag{25}$$

follows (Friedman 2001, Bühlmann and Hothorn 2007).

Friedman *et al.* (2000) show how to derive the population minimizer of the negative binomial log-likelihood in equation (12). For notational convenience, we encode the response by $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$. The likelihood in (12) can then be written as

$$L(\tilde{Y}, F) = \log\left(1 + e^{-\tilde{Y}F}\right). \tag{26}$$

In analogy to our previous definition, we set $p(\mathbf{x}) := \Pr[\tilde{Y} = 1\,|\,\mathbf{X} = \mathbf{x}]$, and hence $1 - p(\mathbf{x}) := \Pr[\tilde{Y} = -1\,|\,\mathbf{X} = \mathbf{x}]$. Dropping the arguments, we have

$$\begin{aligned}\mathbb{E}_{\tilde{Y}}[L\,|\,\mathbf{x}] &= \mathbb{E}_{\tilde{Y}}\left[\log(1 + e^{\tilde{Y}F})\,\middle|\,\mathbf{x}\right] \\ &= p\log(1 + e^{-F}) + (1-p)\log(1 + e^{F}).\end{aligned} \tag{27}$$

The partial derivative with respect to F is

$$\mathbb{E}_{\tilde{Y}}\left[\log(1 + e^{\tilde{Y}F})\,\middle|\,\mathbf{x}\right] = -p\,\frac{e^{-F}}{1 + e^{-F}} + (1-p)\,\frac{e^{F}}{1 + e^{F}}. \tag{28}$$

Setting to zero and solving for F, we obtain the population minimizer

$$F^*(\mathbf{x}) = \log\left[\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right]. \tag{29}$$

Notice that Friedman *et al.* (2000) and Bühlmann and Hothorn (2007) use a slightly different parameterization, namely setting F equal to *one half* of the logit-transform, such as to have the population minimizer equal to the one for the exponential loss criterion. The population minimizers in (23), (25) and (29) imply that the initial function estimates in step 1 of the FGD algorithm (Appendix) must be set to $F^*(\cdot) \equiv \mathrm{median}(Y)$ for the L_1-loss, to $F^*(\cdot) \equiv \bar{Y}$ for the L_2-loss, and to $F^*(\cdot) \equiv \log[\hat{p}/(1-\hat{p})]$ for the negative binomial log-likelihood loss.

SI.2.2 Negative gradient

To calculate the negative gradient vector $(U_1, \ldots, U_n)$ in step 2 of the FGD algorithm (Appendix), we need the partial derivative of the loss function with respect to the target function F. Any element U_i is obtained as this partial derivative evaluated at the previous function estimate $\hat{F}^{[m-1]}(\mathbf{x}_i)$. Formally,

$$U_i = -\frac{\partial}{\partial F} L(Y_i, F)\Big|_{F=\hat{F}^{[m-1]}(\mathbf{X}_i)}. \tag{30}$$

For the L_1-loss in (9), we have

$$-\frac{\partial}{\partial F}\Big[|Y_i - F|\Big] = \frac{Y_i - F}{|Y_i - F|} = \mathrm{sgn}(Y_i - F), \tag{31}$$

which implies the negative gradient component

$$U_i = \mathrm{sgn}\Big[Y_i - \hat{F}^{[m-1]}(\mathbf{X}_i)\Big] \tag{32}$$

in step 2 of the FGD algorithm (*cf.* Friedman 2001).

For the L_2-loss in (10),

$$-\frac{\partial}{\partial F}\left[\frac{1}{2}(Y_i - F)^2\right] = Y_i - F, \tag{33}$$

which amounts to

$$U_i = Y_i - \hat{F}^{[m-1]}(\mathbf{X}_i) \tag{34}$$

in step 2 of the FGD algorithm (*cf.* Friedman 2001, Bühlmann and Hothorn 2007).

Last, for the negative binomial log-likelihood we again use $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$ and find

$$-\frac{\partial}{\partial F} L(\tilde{Y}_i, F) = -\frac{\partial}{\partial F}\log\left(1 + e^{-\tilde{Y}_i F}\right) = \frac{\tilde{Y}_i\, e^{-\tilde{Y}_i F}}{1 + e^{-\tilde{Y}_i F}}. \tag{35}$$

This leads to the negative gradient component

$$U_i = \frac{\tilde{Y}_i\, e^{-\tilde{Y}_i \hat{F}^{[m-1]}(\mathbf{X}_i)}}{1 + e^{-\tilde{Y}_i \hat{F}^{[m-1]}(\mathbf{X}_i)}} \tag{36}$$

in step 2 of the FGD algorithm.

SI Supporting Information: URLs to Supporting Files

File S2
Census population sizes of Alpine ibex demes in the Swiss Alps

File S2 is available for download as a PDF file at http://pub.ist.ac.at/~saeschbacher/phd_e-sources/[to be found in subsection 'Chapter 3'].

File S3
Numbers of Alpine ibex transferred between demes by humans

File S3 is available for download as a PDF file at http://pub.ist.ac.at/~saeschbacher/phd_e-sources/[to be found in subsection 'Chapter 3'].

Literature Cited

Bühlmann, P. and T. Hothorn, 2007 Boosting algorithms: Regularization, prediction and model fitting. Stat. Sci. **22**: 477–505.

Friedman, J., T. Hastie, and R. Tibshirani, 2000 Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting. Ann. Stat. **28**: 337–374.

Friedman, J. H., 2001 Greedy function approximation: A gradient boosting machine. Ann. Stat. **29**: 1189–1232.

Loison, A., C. Toïgo, J. Appolinaire, and J. Michallet, 2002 Demographic processes in colonizing populations of isard (*Rupicapra pyrenaica*) and ibex (*Capra ibex*). J. Zool. **256**: 199–205.

Nievergelt, B., 1966 *Der Alpensteinbock (*Capra ibex L.*) in seinem Lebensraum. Ein ökologischer Vergleich*. Mammalia depicta, Verlag Paul Parey, Hamburg, Berlin.

Stuwe, M. and C. Grodinsky, 1987 Reproductive biology of captive Alpine ibex (*Capra i. ibex*). Zoo Biol. **6**: 331–339.

Toïgo, C., J. M. Gaillard, M. Festa-Bianchet, E. Largo, J. Michallet, and D. Maillard, 2007 Sex- and age-specific survival of the highly dimorphic Alpine ibex: evidence for a conservative life-history tactic. J. Anim. Ecol. **76**: 679–686.

Toïgo, C., J. M. Gaillard, D. Gauthier, I. Girard, J. P. Martinot, and J. Michallet, 2002 Female reproductive success and costs in an alpine capital breeder under contrasting environments. Écoscience **9**: 427–433.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182**: 1207–1218.

A novel approach for choosing summary statistics in approximate Bayesian computation

– Online supporting information –

Simon Aeschbacher[*,§], Mark A. Beaumont[**], Andreas Futschik[§§]

August 24, 2012

[*] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, [§] IST Austria (Institute of Science and Technology Austria), 3400 Klosterneuburg, Austria, [**] Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TW, United Kingdom, and [§§] Institute of Statistics and Decision Support Systems, University of Vienna, 1010 Vienna, Austria

SI Supporting Information: Overview

This file contains

- Additional Tables (Table S1)
- Additional Figures (Figures S1–S15)
- Additional Methods
- URLs to two Supporting Files (Files S2 and S3)

SI Supporting Information: Additional Tables

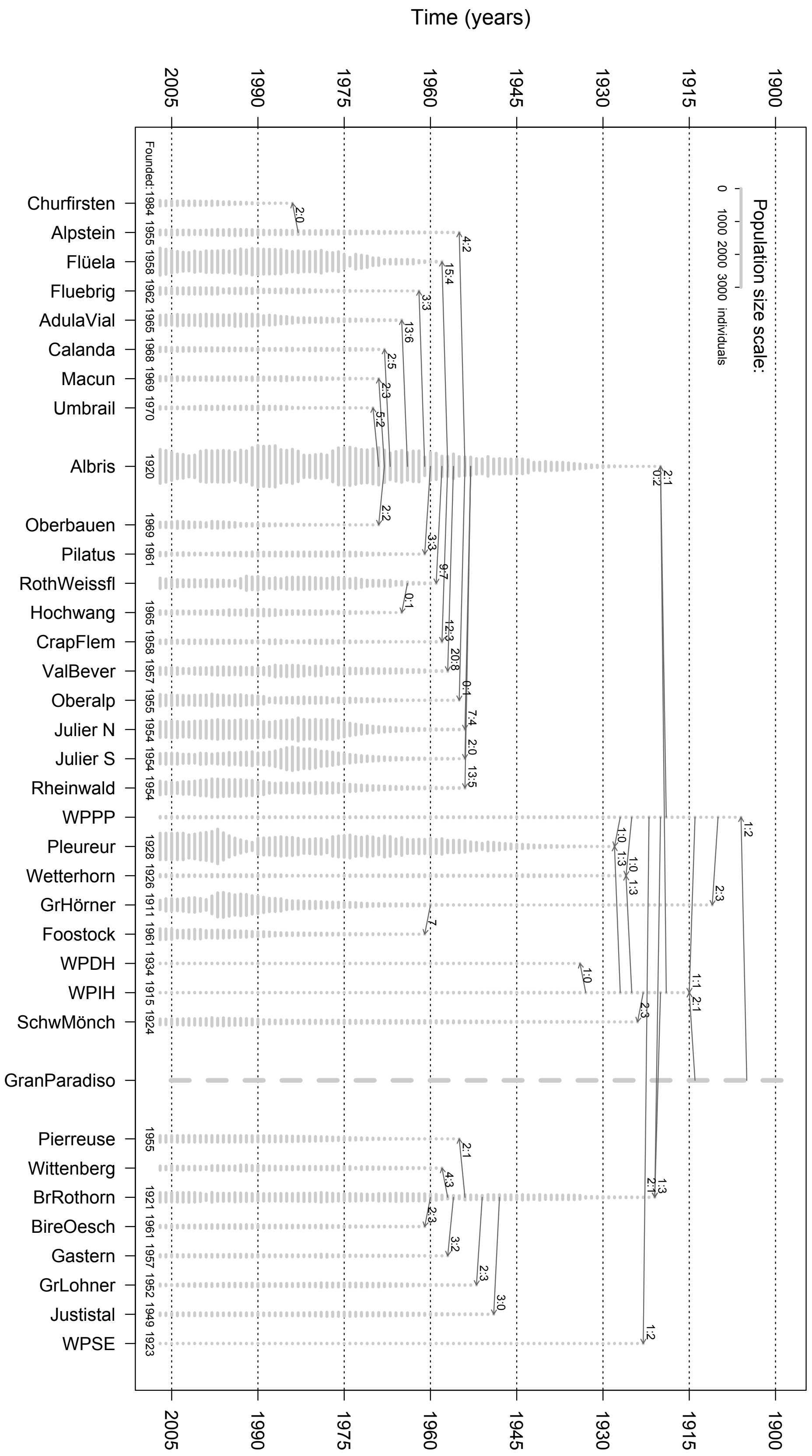
Table S1 Deme names, deme numbers and sampling sizes in the Alpine ibex data set

				Genetic sample size[c]		
Deme name	Deme number[a]	Short name	Internal number[b]	Males	Females	Total
Adula Vial	1	AdulaVial	100	21	16	37
Albris	2	Albris	101	28	33	61
Alpstein	3	Alpstein	102	12	18	30
Bire-Oeschinen	4	BireOesch	103	16	2	18
Brienzer Rothorn	5	BrRothorn	104	21	18	39
Calanda	6	Calanda	105	15	16	31
Churfirsten	7	Churfirsten	106	11	13	24
Crap da Flem	8	CrapFlem	107	16	11	27
Fluebrig	9	Fluebrig	108	17	15	32
Flüela	10	Flüela	109	37	38	75
Foostock	11	Foostock	110	9	18	27
Gastern	12	Gastern	111	5	6	11
Graue Hörner	13	GrHörner	112	21	26	47
Gross Lohner	14	GrLohner	113	15	7	22
Hochwang	15	Hochwang	114	14	14	28
Julier Nord	16	Julier N	115	12	11	23
Julier Süd	17	Julier S	116	12	11	23
Justistal	18	Justistal	117	15	4	19
Macun	19	Macun	118	12	10	22
Oberalp-Frisal	20	Oberalp	134	25	19	44
Oberbauenstock	21	Oberbauen	119	18	12	30
Pilatus	22	Pilatus	120	15	2	17
Mont Pleureur	23	Pleureur	121	22	7	29
Safien-Rheinwald	24	Rheinwald	122	22	13	35
Rothorn-Weissfluh	25	RothWeissfl	123	16	13	29
Schwarzmönch	26	SchwMönch	124	15	17	32
Umbrail	27	Umbrail	125	15	14	29
Val Bever	28	ValBever	126	20	12	32
Wetterhorn	29	Wetterhorn	127	9	10	19
Wittenberg	30	Wittenberg	128	15	6	21
Pierreuse-Gummfluh	31	Pierreuse	133	20	21	41
Wildpark Dählhölzli	32	WPDH	129	0	0	0
Wildpark Interlaken	33	WPIH	130	0	0	0
Wildpark St. Gallen	34	WPPP	131	0	0	0
Wildpark Seiler	35	WPSE	132	0	0	0

[a] As used in main text and Figure 1.
[b] As used in scripts and Supporting Files S2 and S3.
[c] The number of individuals from which genetic samples were taken, both in reality and in the simulations.

SI Supporting Information: Additional Figures

Figure S1 ***(facing page)*** Genealogy and demography of Alpine ibex demes analyzed in this study. Time goes from top to bottom, starting in the year 1900 and ending in 2007. Horizontal gray bars represent the known census sizes (Supporting File S2 *census sizes*) and arrows show the founder events by which demes were established. The numbers of males and females transferred are given close to the arrow head (males:females; for *Foostock*, the sex of the founders is unknown and only the total number of founders is given). Most demes received further individuals after the initial founder event, but these numbers are not shown here (see Supporting File S3 *transfers*). The deme ancestral to all other demes, *GranParadiso*, is shown as a vertical dashed line; its deme size is not known. See also Table S1 for the full deme names and Figure 1 for the geographical location of demes.


Time (years)
2005
1990
1975
1960
1945
1930
1915
1900
Population size scale:
0 1000 2000 3000 individuals
Founded:
Churfirsten 1984
Alpstein 1955
Flüela 1958
Fluebrig 1962
AdulaVial 1965
Calanda 1968
Macun 1969
Umbrail 1970
Albris 1920
Oberbauen 1969
Pilatus 1961
RothWeissfl
Hochwang 1965
CrapFlem 1958
ValBever 1957
Oberalp 1955
Julier N 1954
Julier S 1954
Rheinwald 1954
WPPP
Pleureur 1928
Wetterhorn 1926
GrHörner 1911
Foostock 1961
WPDH 1934
WPIH 1915
SchwMönch 1924
GranParadiso
Pierreuse 1955
Wittenberg
BrRothorn 1921
BireOesch 1961
Gastern 1957
GrLohner 1952
Justistal 1949
WPSE 1923
2:0
4:2
15:4
3:3
13:6
2:5
2:3
5:2
2:1
0:2
2:2
3:3
9:7
0:1
12:3
20:8
0:1
7:4
2:0
13:5
1:2
1:0
1:3
1:0
1:3
2:3
7
1:0
1:1
2:1
2:3
2:1
4:3
1:3
2:1
2:3
3:2
2:3
3:0
1:2

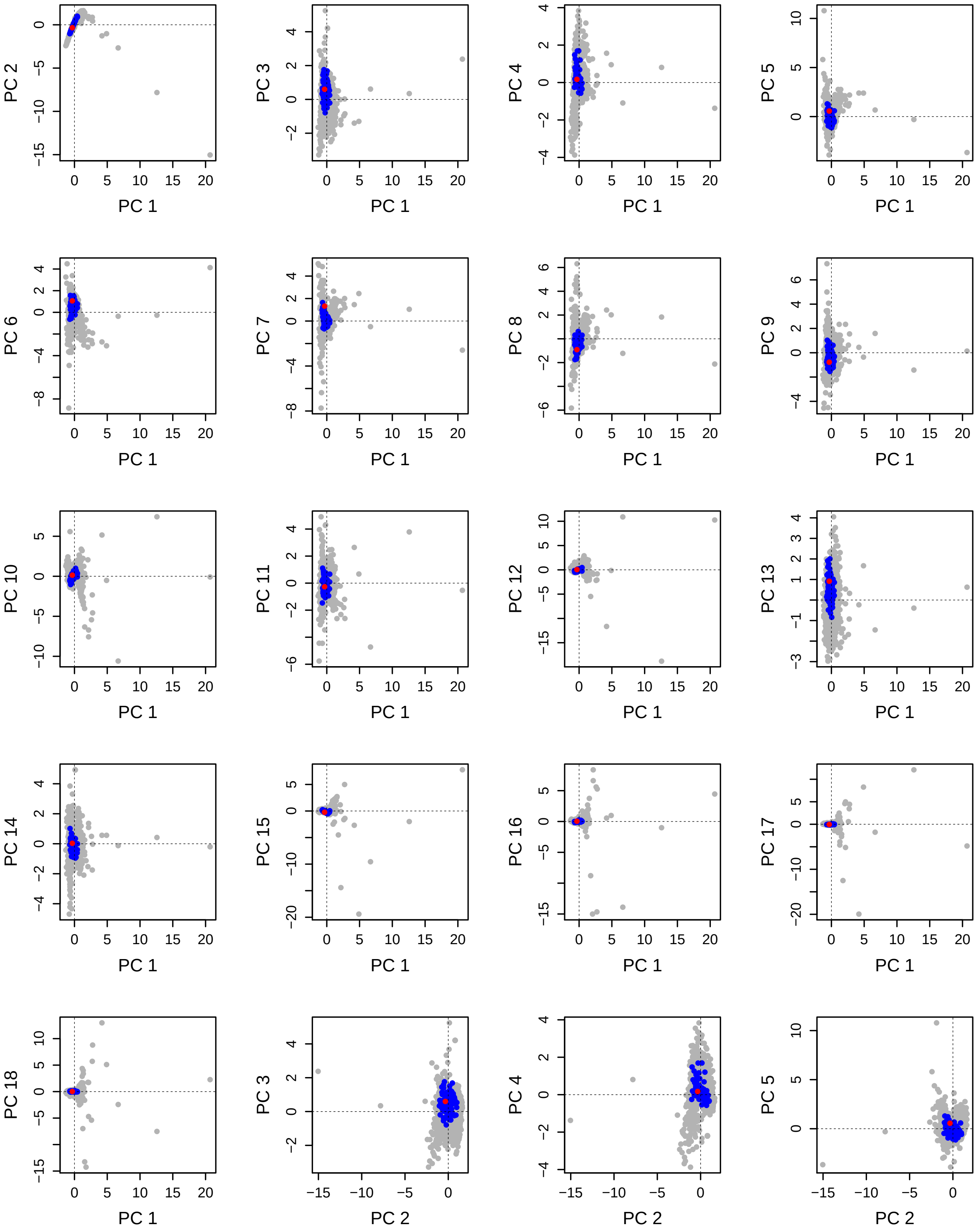


Figure S2 *Continued on next page*

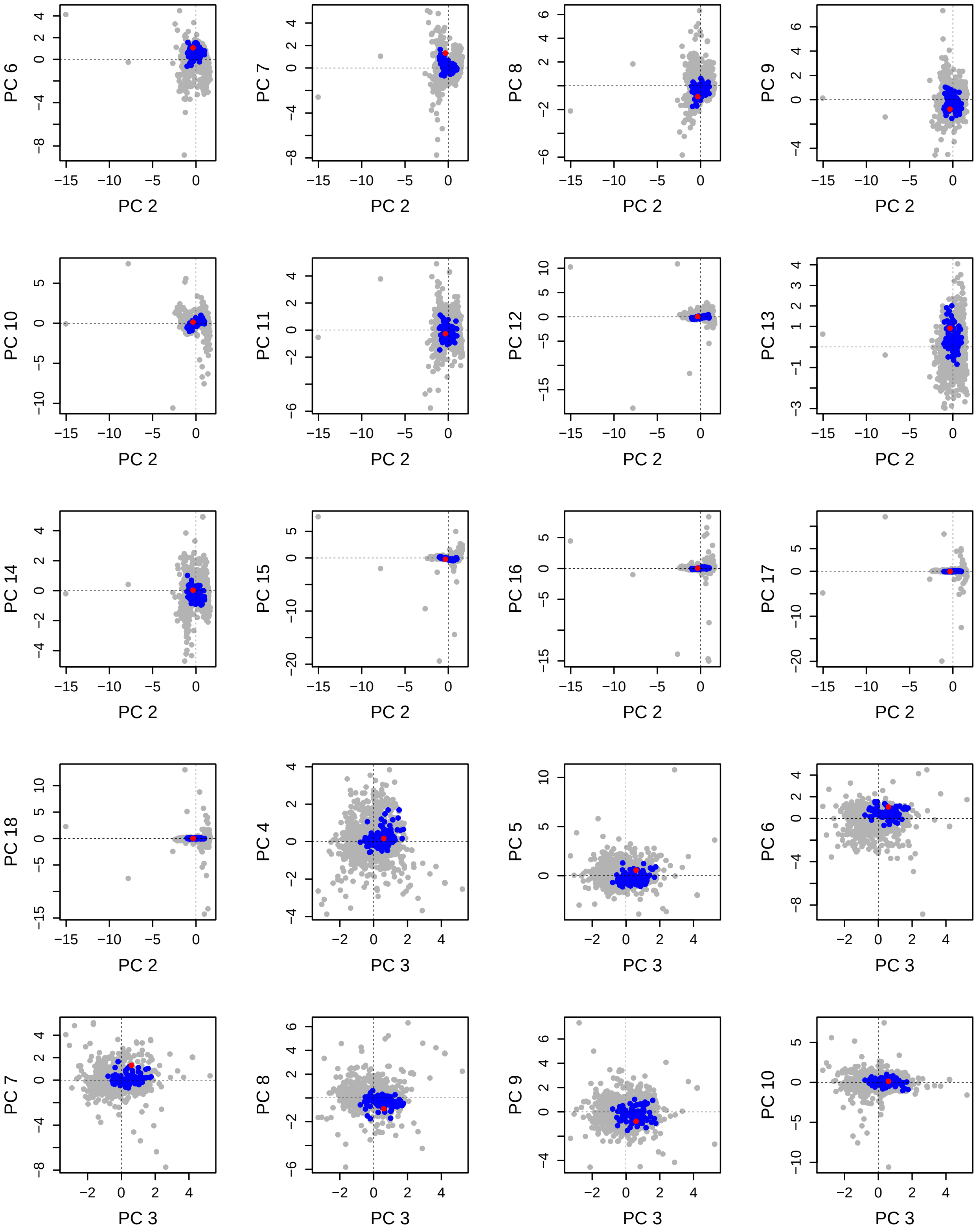


Figure S2 *Continued on next page*

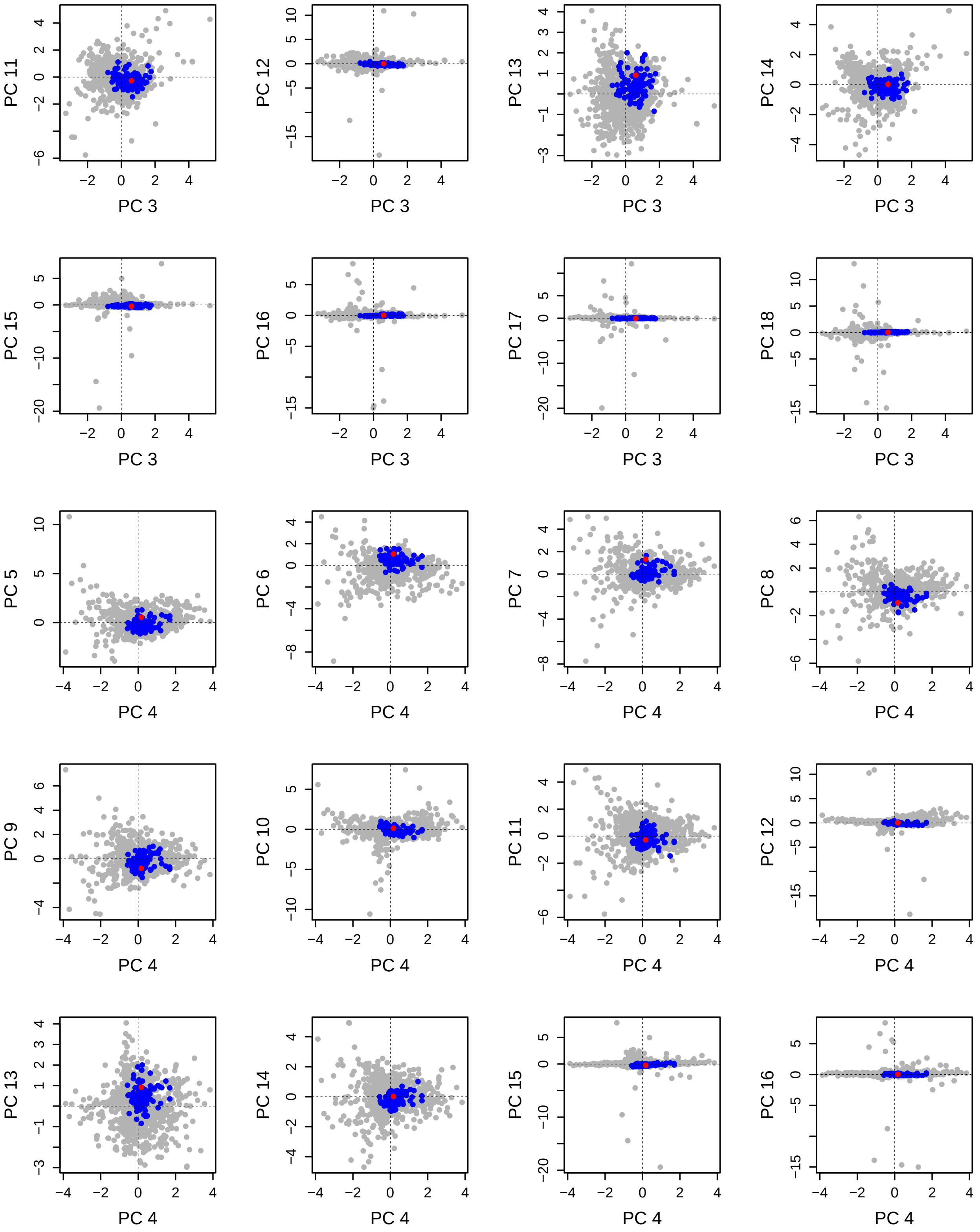


Figure S2 *Continued on next page*

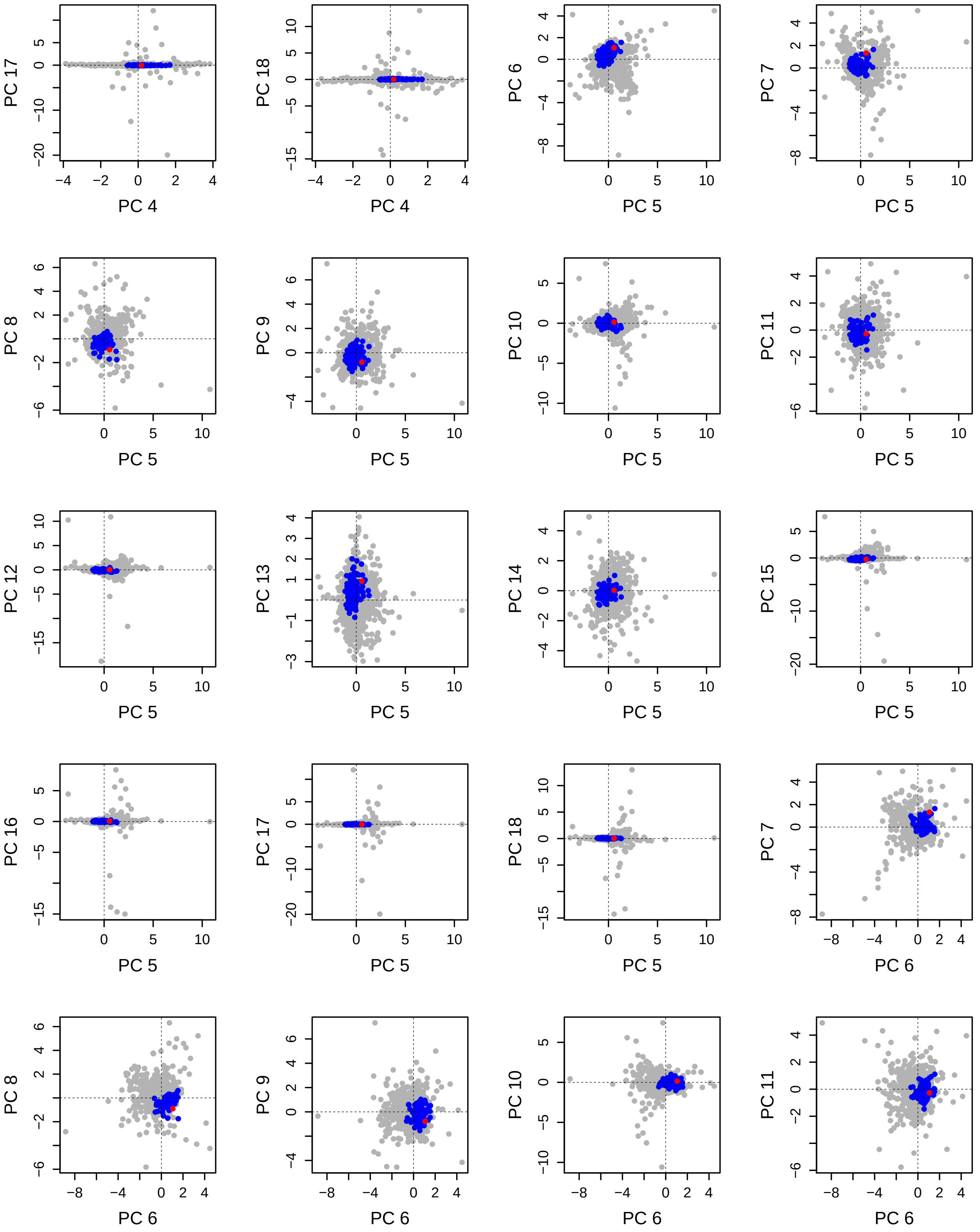


Figure S2 *Continued on next page*

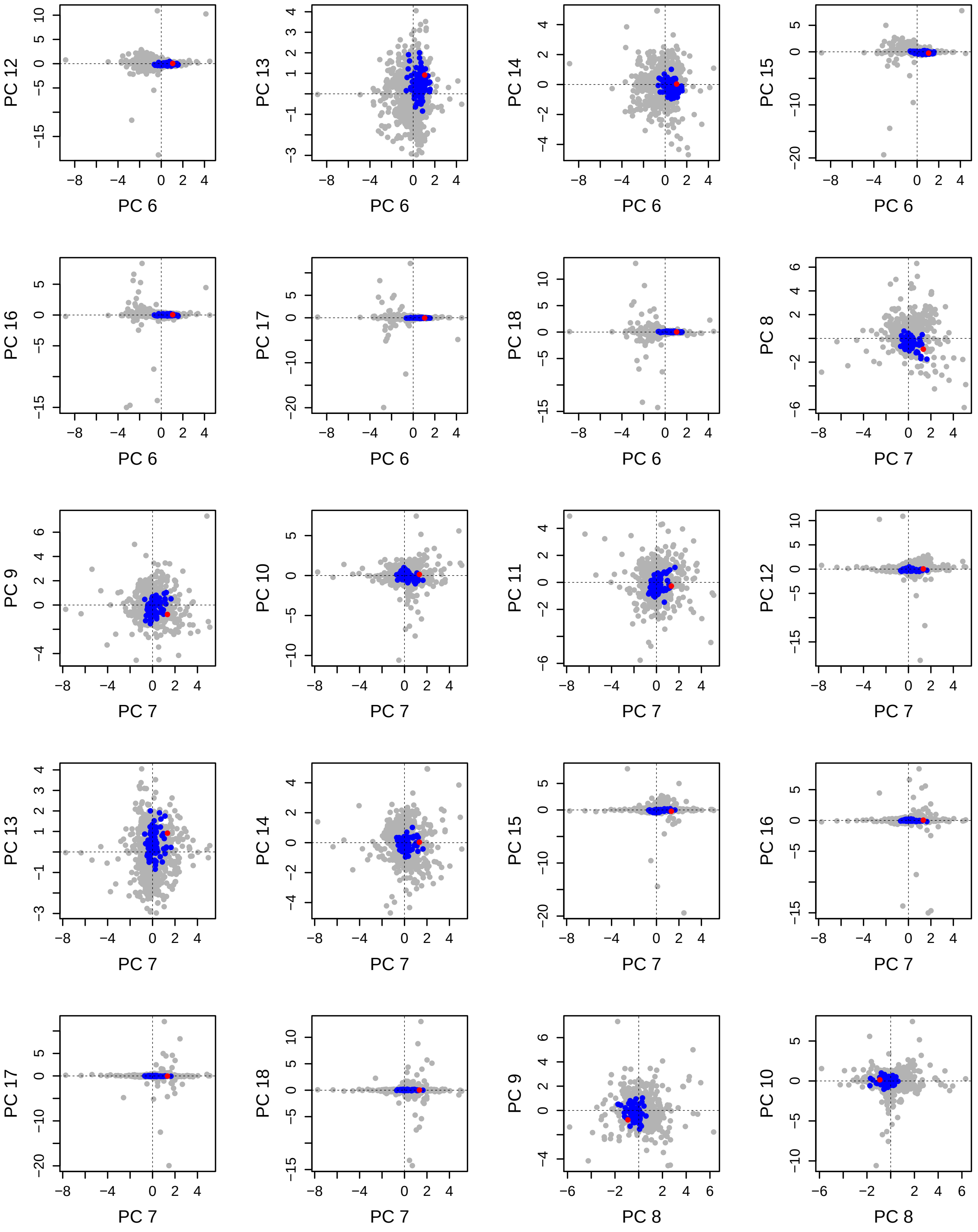


Figure S2 *Continued on next page*

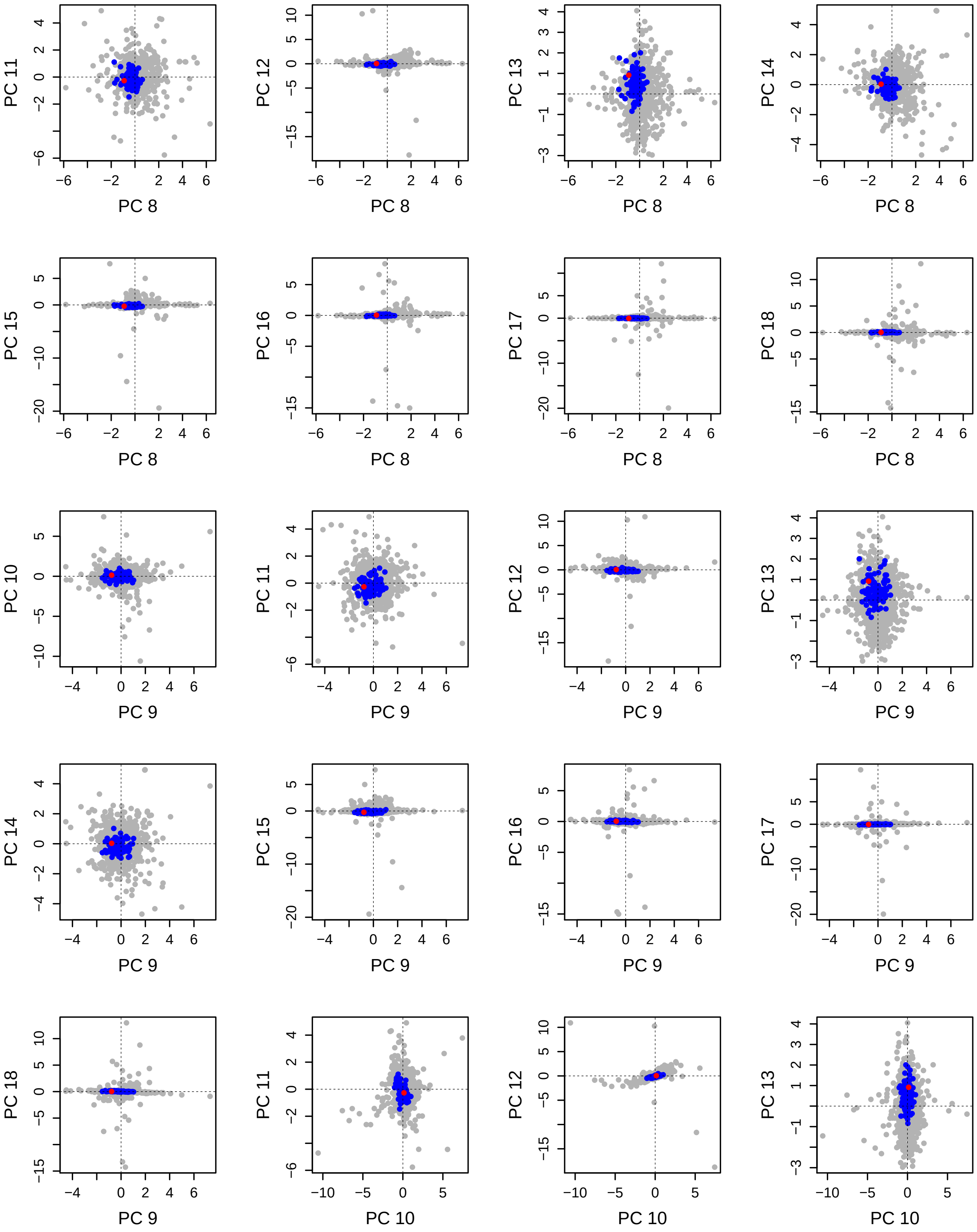


Figure S2 *Continued on next page*

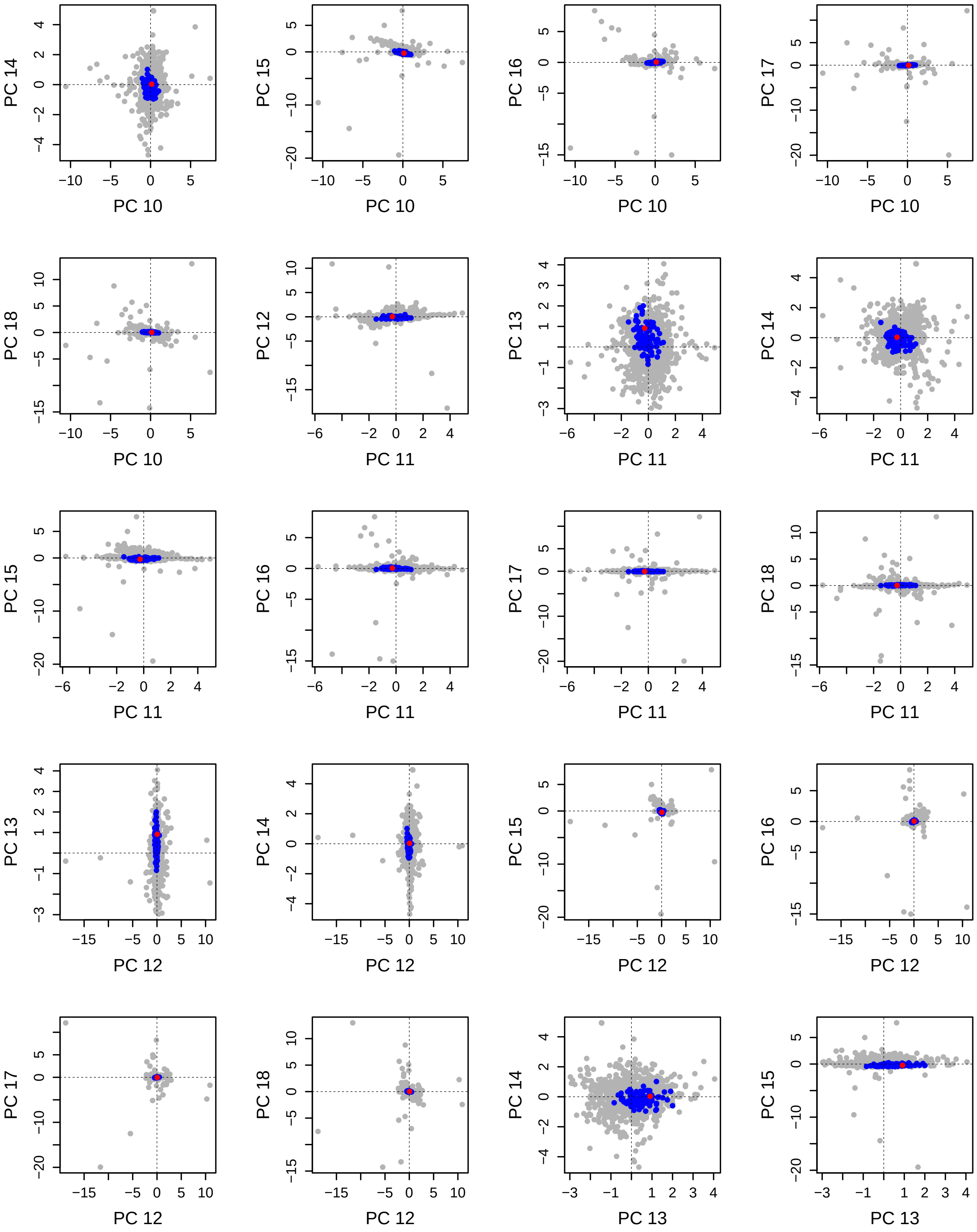


Figure S2 *Continued on next page*

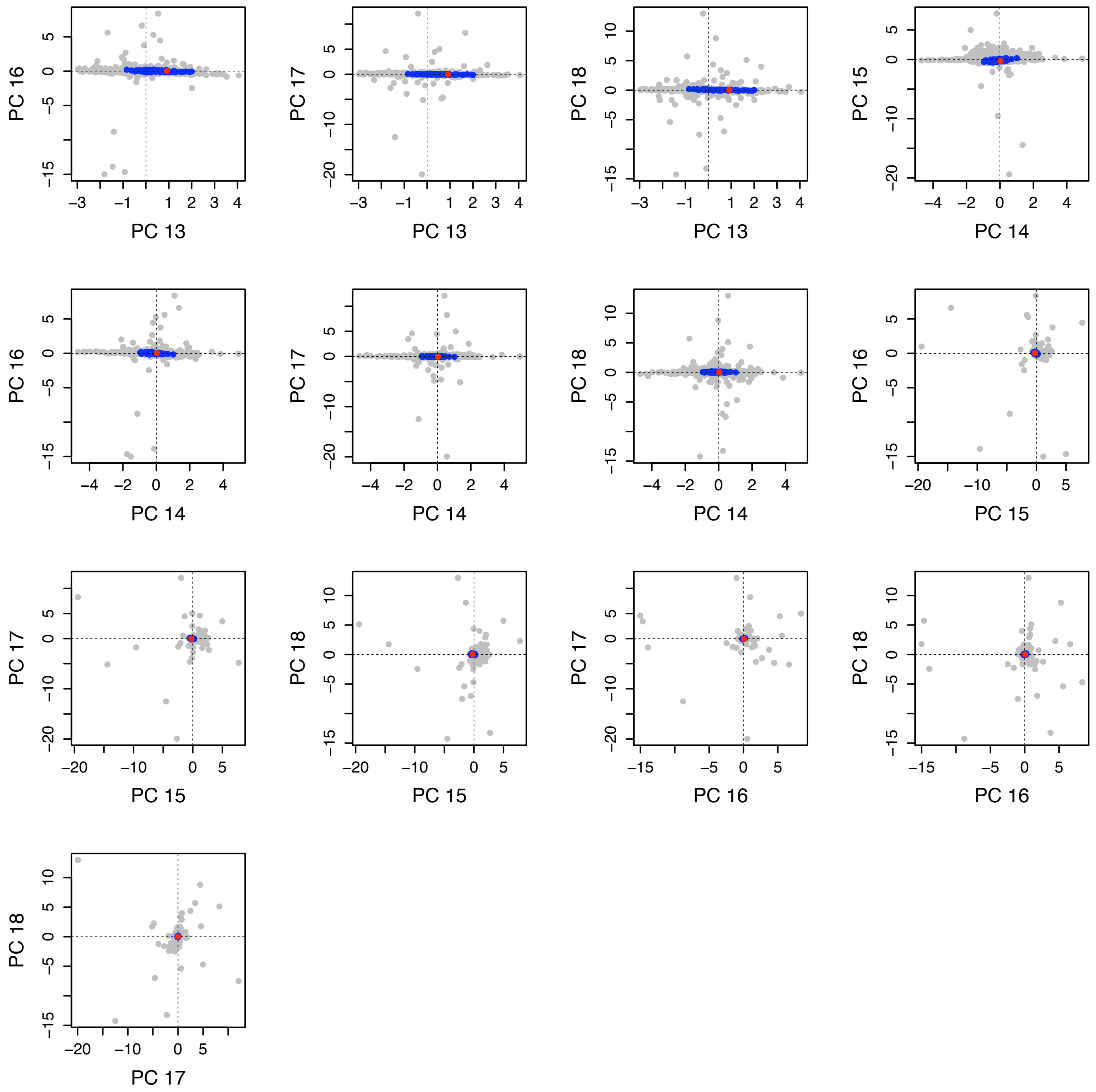


Figure S2 *Continued from previous page.* Pairwise prior predictive distribution of PC-rotated summary statistics. Gray points represent $N = 1000$ simulations with parameter values drawn from the prior. The true value from the ibex data set is shown as a red dot. The fact that it is always embedded in the cloud of gray points means that the model and prior distributions are well specified. The $n' = 100$ points with smallest Euclidean distance from the observation are shown in blue. Those represent simulations used as training data sets for the *local* choice of summary statistics (see main text). In the main study, we used $N = 10^6$ and $n' = 1000$; smaller numbers are used here for illustration of the principle.

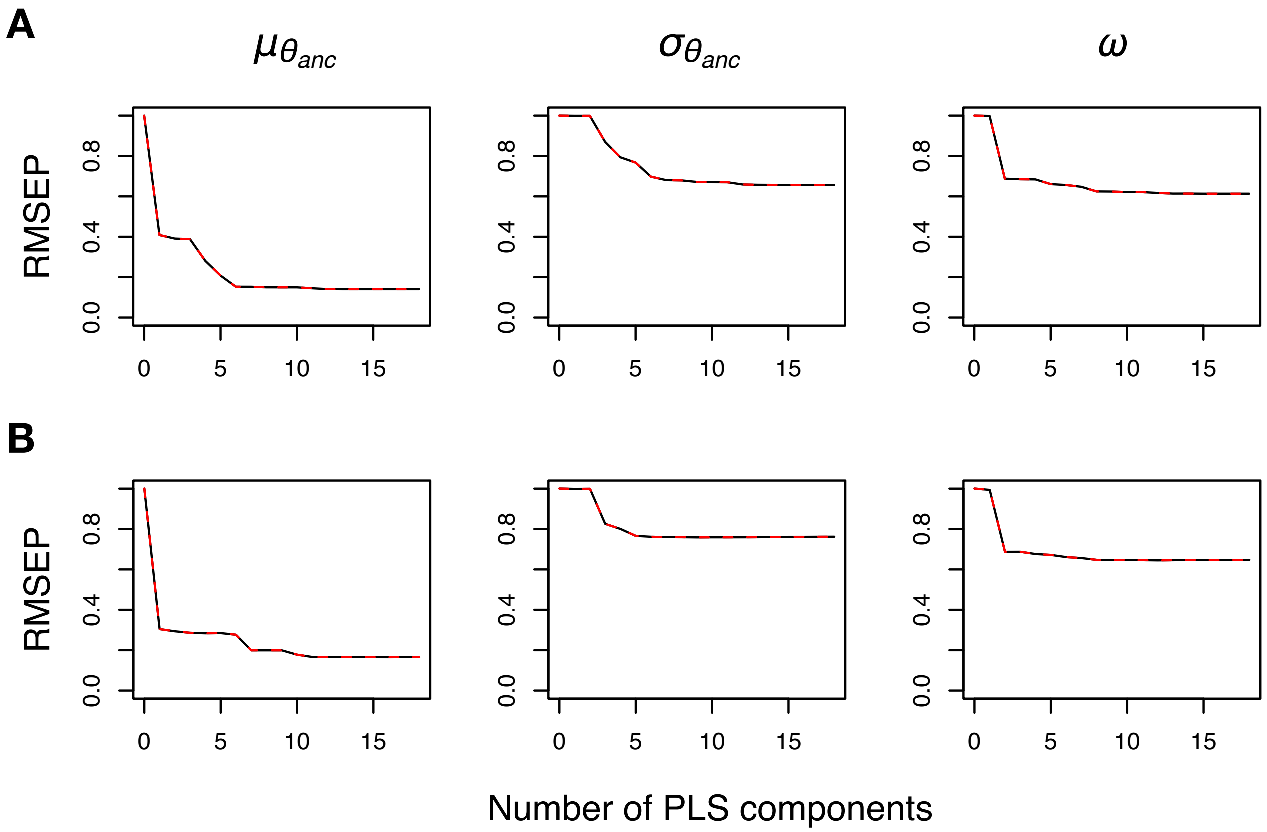


Figure S3 Root mean squared error of prediction (RMSEP) for PLS regression as a function of the number of PLS components used. As suggested by (Wegmann *et al.* 2009), we chose the number of PLS components to be kept as summary statistics based on these plots. The RMSEP was obtained via leave-one-out cross-validation. (A) Global and (B) local choice of summary statistics via PLS (see main text). In (B), the observation from the ibex data set was used as the center. In both cases, we decided to keep the first ten components as summary statistics.

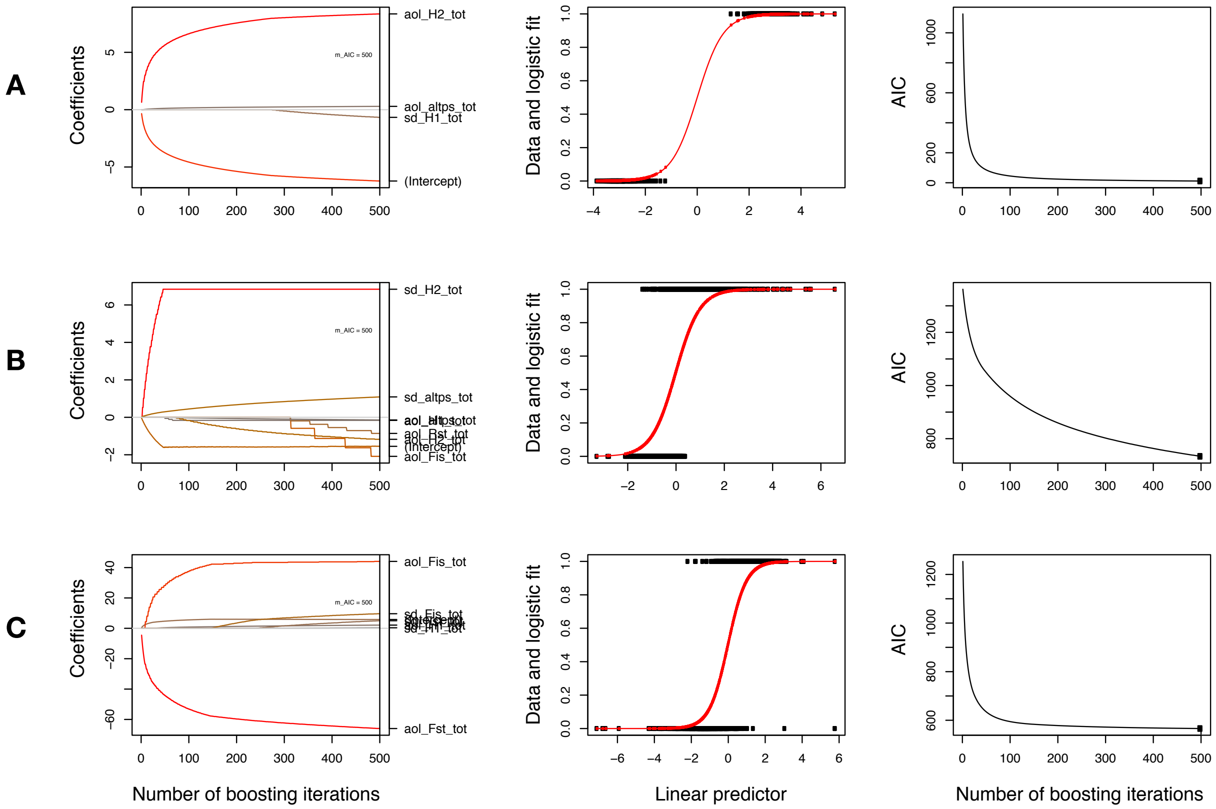


Figure S4 Choice of summary statistics via LogitBoost for the three parameters $\mu_{\theta_{\rm anc}}$ (A), $\sigma_{\theta_{\rm anc}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Middle column: Binary parameter class variable (Y, black) and logistic fit to the probability $\Pr[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ (red), as a function of the linear predictor. Right column: Quality of fit in terms of AIC as a function of the number of iterations m. The thick black line marks the m_{stop} chosen. In the cases shown here, no minimum AIC was found for $m < 500$.

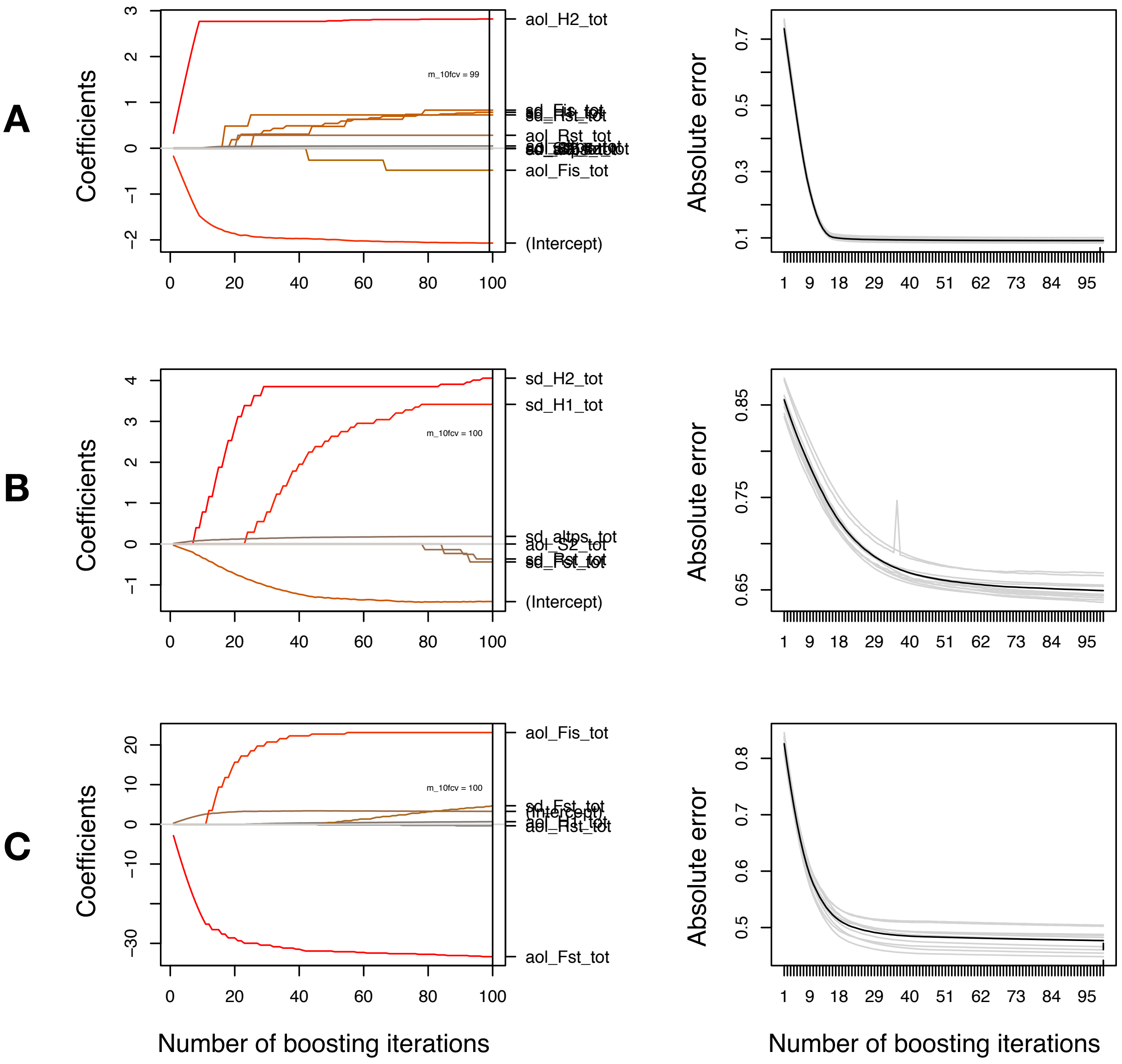


Figure S5 Choice of summary statistics via L_1Boosting for the three parameters $\mu_{\theta_{\mathrm{anc}}}$ (A), $\sigma_{\theta_{\mathrm{anc}}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Right column: Quality of fit in terms of the bootstrapping error, as a function of the number of iterations m. The dashed vertical line marks the m_{stop} chosen. In the cases shown here, no minimum absolute error was found for $m < 100$.

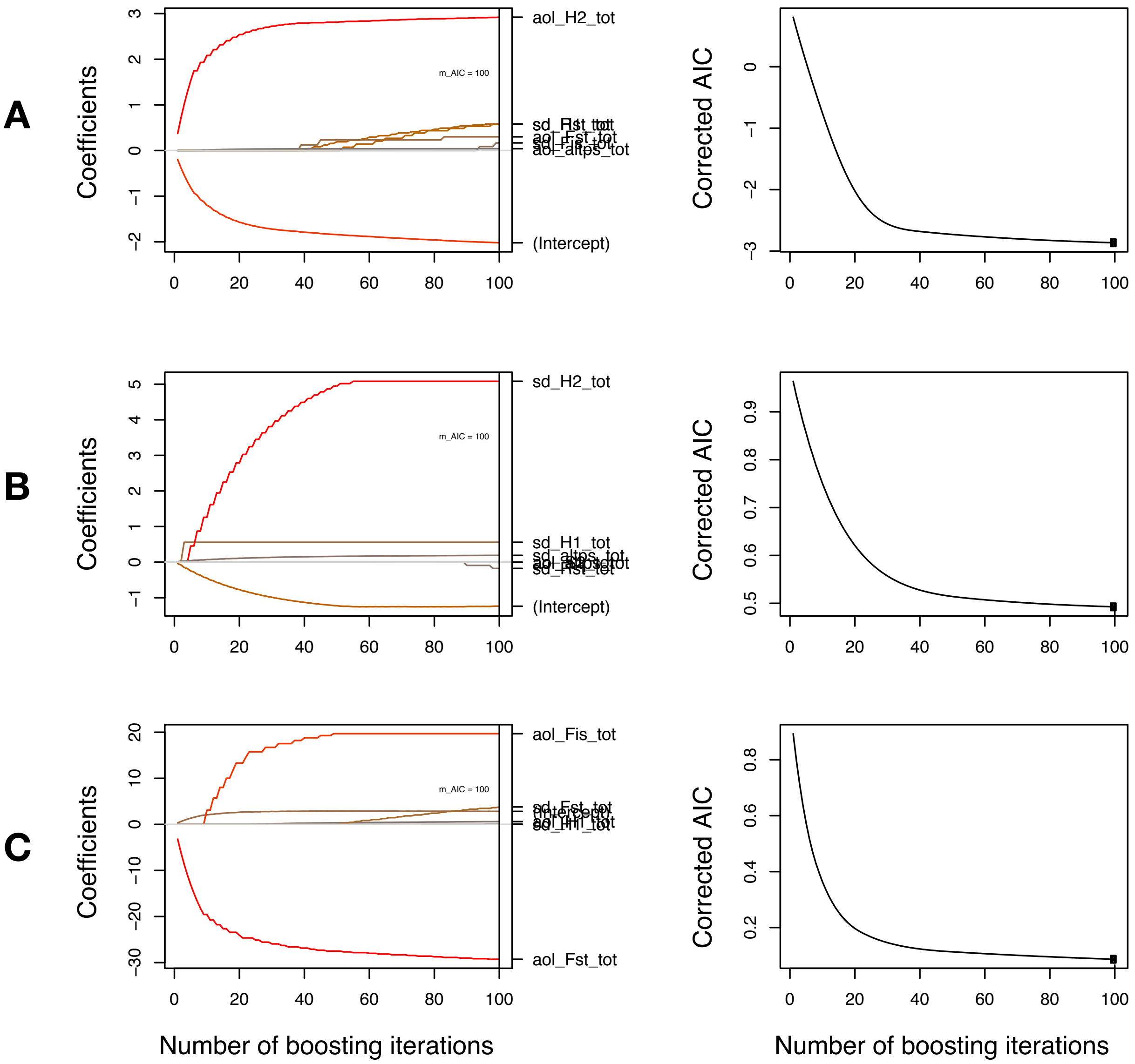


Figure S6 Choice of summary statistics via L_2Boosting for the three parameters $\mu_{\theta_{\mathrm{anc}}}$ (A), $\sigma_{\theta_{\mathrm{anc}}}$ (B) and ω (C). Left column: Boosted coefficients $\lambda^{[m]}$ as a function of the number of iterations m. Right column: Quality of fit in terms of the corrected AIC as a function of the number of iterations m. The thick black line marks the m_{stop} chosen. In the cases shown here, no minimum absolute error was found for $m < 100$.

Figure S7 ***(facing page)*** Posterior distributions inferred for six random test data sets with acceptance rate $\epsilon = 0.01$. Methods are as described in the main text. True values are given by a dashed vertical line, prior distributions in blue (*cf.* Table 1).

ε = 0.01

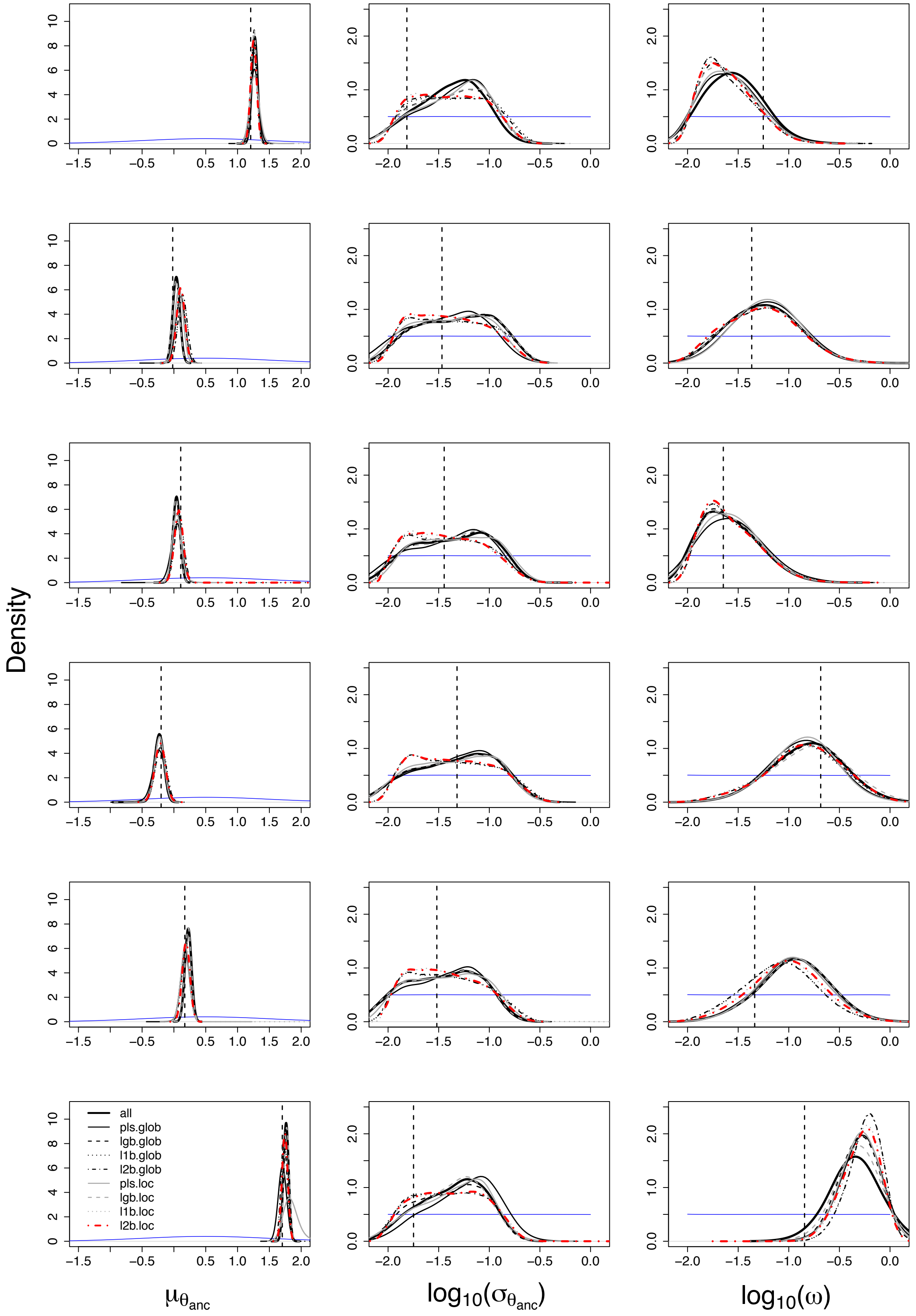

Density
all
pls.glob
lgb.glob
l1b.glob
l2b.glob
pls.loc
lgb.loc
l1b.loc
l2b.loc
$\mu_{\theta_{anc}}$
$\log_{10}(\sigma_{\theta_{anc}})$
$\log_{10}(\omega)$
−1.5
−0.5
0.5
1.0
1.5
2.0
−2.0
−1.5
−1.0
−0.5
0.0
0
2
4
6
8
10
0.0
1.0
2.0

Figure S8 ***(facing page)*** Coverage property of posterior distributions inferred with different choices of summary statistics on a global scale. Histograms show the distribution across 500 independent test estimations of the posterior probabilities of the true parameter values. The distribution is expected to be uniform (Wegmann *et al.* 2009). Left-skewed or right-skewed distributions indicate that the parameter is on average over- or underestimated, respectively. Peaked or U-shaped distributions result from posterior distributions that are too wide or too narrow, respectively. Non-uniform distributions are shaded in gray (p-values from a Kolmogorov-Smirnov test on top are without correction for multiple testing; see main text).

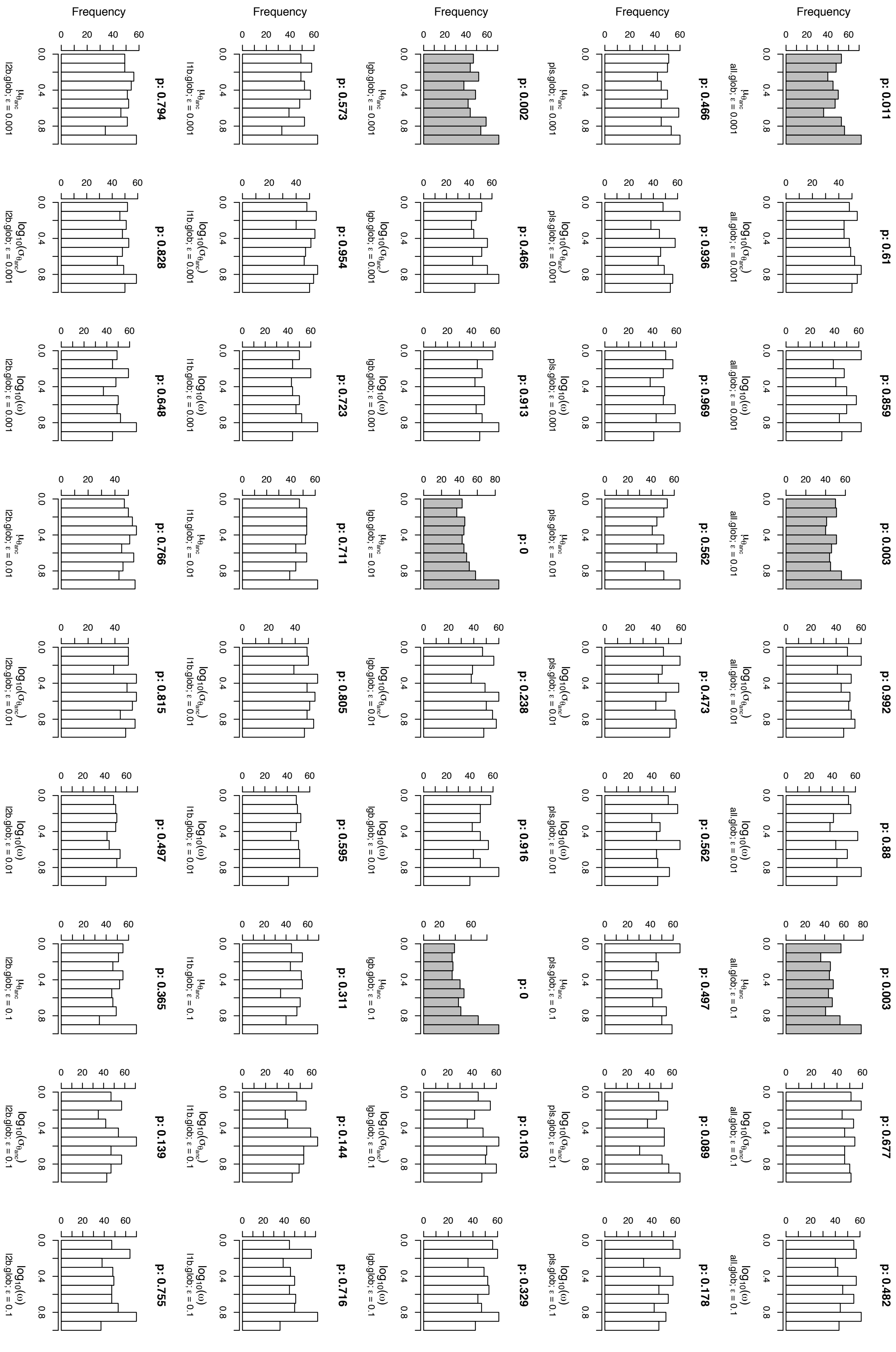

Frequency
p: 0.011
$\mu_{\theta_{anc}}$ all.glob; ε = 0.001
p: 0.61
$\log_{10}(\sigma_{\theta_{anc}})$ all.glob; ε = 0.001
p: 0.859
$\log_{10}(\omega)$ all.glob; ε = 0.001
p: 0.003
$\mu_{\theta_{anc}}$ all.glob; ε = 0.01
p: 0.992
$\log_{10}(\sigma_{\theta_{anc}})$ all.glob; ε = 0.01
p: 0.88
$\log_{10}(\omega)$ all.glob; ε = 0.01
p: 0.003
$\mu_{\theta_{anc}}$ all.glob; ε = 0.1
p: 0.677
$\log_{10}(\sigma_{\theta_{anc}})$ all.glob; ε = 0.1
p: 0.482
$\log_{10}(\omega)$ all.glob; ε = 0.1
Frequency
p: 0.466
$\mu_{\theta_{anc}}$ pls.glob; ε = 0.001
p: 0.936
$\log_{10}(\sigma_{\theta_{anc}})$ pls.glob; ε = 0.001
p: 0.969
$\log_{10}(\omega)$ pls.glob; ε = 0.001
p: 0.562
$\mu_{\theta_{anc}}$ pls.glob; ε = 0.01
p: 0.473
$\log_{10}(\sigma_{\theta_{anc}})$ pls.glob; ε = 0.01
p: 0.562
$\log_{10}(\omega)$ pls.glob; ε = 0.01
p: 0.497
$\mu_{\theta_{anc}}$ pls.glob; ε = 0.1
p: 0.089
$\log_{10}(\sigma_{\theta_{anc}})$ pls.glob; ε = 0.1
p: 0.178
$\log_{10}(\omega)$ pls.glob; ε = 0.1
Frequency
p: 0.002
$\mu_{\theta_{anc}}$ lgb.glob; ε = 0.001
p: 0.466
$\log_{10}(\sigma_{\theta_{anc}})$ lgb.glob; ε = 0.001
p: 0.913
$\log_{10}(\omega)$ lgb.glob; ε = 0.001
p: 0
$\mu_{\theta_{anc}}$ lgb.glob; ε = 0.01
p: 0.238
$\log_{10}(\sigma_{\theta_{anc}})$ lgb.glob; ε = 0.01
p: 0.916
$\log_{10}(\omega)$ lgb.glob; ε = 0.01
p: 0
$\mu_{\theta_{anc}}$ lgb.glob; ε = 0.1
p: 0.103
$\log_{10}(\sigma_{\theta_{anc}})$ lgb.glob; ε = 0.1
p: 0.329
$\log_{10}(\omega)$ lgb.glob; ε = 0.1
Frequency
p: 0.573
$\mu_{\theta_{anc}}$ l1b.glob; ε = 0.001
p: 0.954
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.glob; ε = 0.001
p: 0.723
$\log_{10}(\omega)$ l1b.glob; ε = 0.001
p: 0.711
$\mu_{\theta_{anc}}$ l1b.glob; ε = 0.01
p: 0.805
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.glob; ε = 0.01
p: 0.595
$\log_{10}(\omega)$ l1b.glob; ε = 0.01
p: 0.311
$\mu_{\theta_{anc}}$ l1b.glob; ε = 0.1
p: 0.144
$\log_{10}(\sigma_{\theta_{anc}})$ l1b.glob; ε = 0.1
p: 0.716
$\log_{10}(\omega)$ l1b.glob; ε = 0.1
Frequency
p: 0.794
$\mu_{\theta_{anc}}$ l2b.glob; ε = 0.001
p: 0.828
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.glob; ε = 0.001
p: 0.648
$\log_{10}(\omega)$ l2b.glob; ε = 0.001
p: 0.766
$\mu_{\theta_{anc}}$ l2b.glob; ε = 0.01
p: 0.815
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.glob; ε = 0.01
p: 0.497
$\log_{10}(\omega)$ l2b.glob; ε = 0.01
p: 0.365
$\mu_{\theta_{anc}}$ l2b.glob; ε = 0.1
p: 0.139
$\log_{10}(\sigma_{\theta_{anc}})$ l2b.glob; ε = 0.1
p: 0.755
$\log_{10}(\omega)$ l2b.glob; ε = 0.1
0.0
0.4
0.8
0
20
40
60
80

Figure S9 ***(facing page)*** Coverage property of posterior distributions inferred with different choices of summary statistics on a local scale. Non-uniform distributions of posterior probabilities are shaded in gray (p-values from a Kolmogorov-Smirnov test on top). Note that the first row here corresponds to the first row in Figure S8. Further details as in Figure S8.

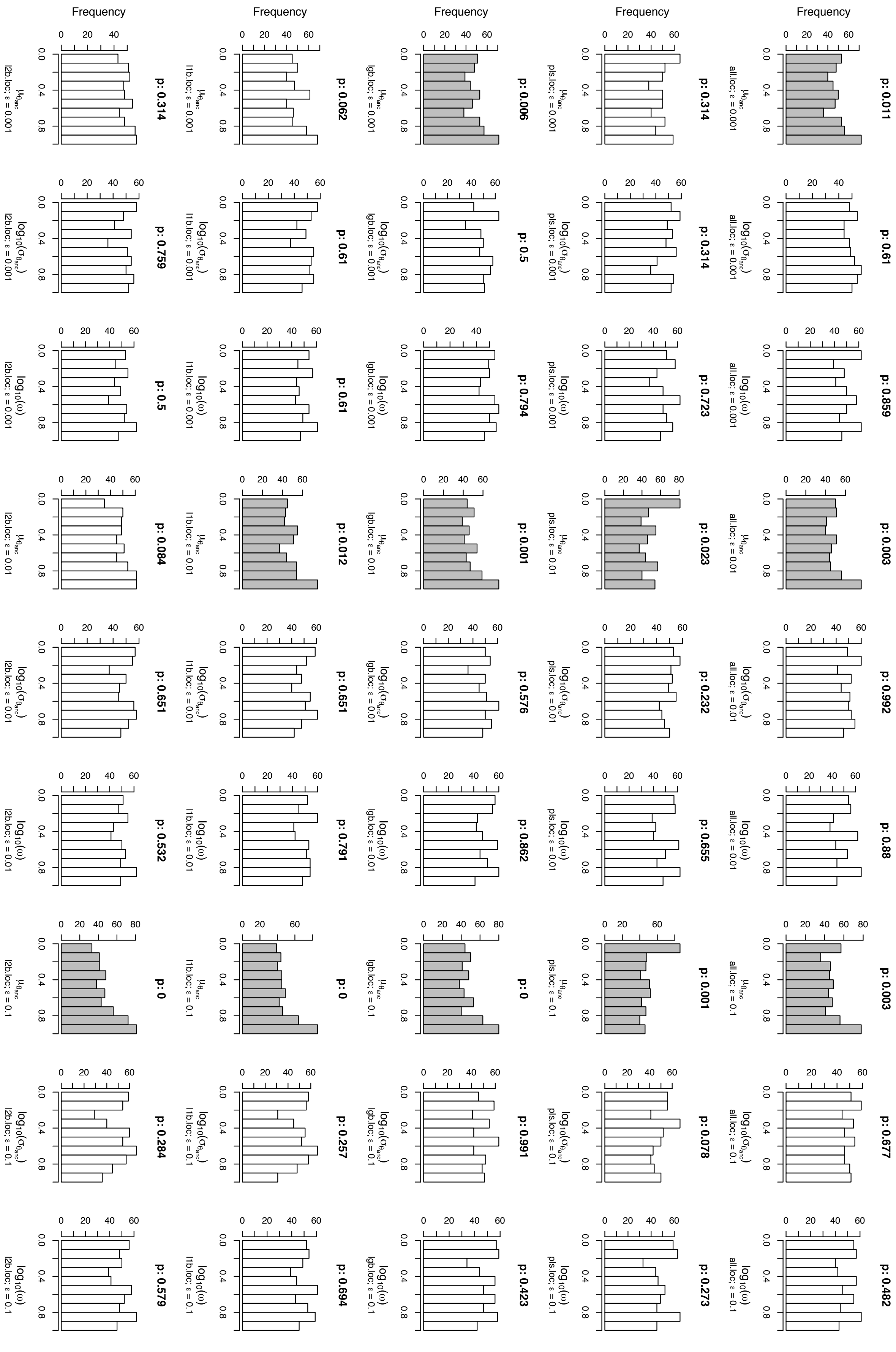
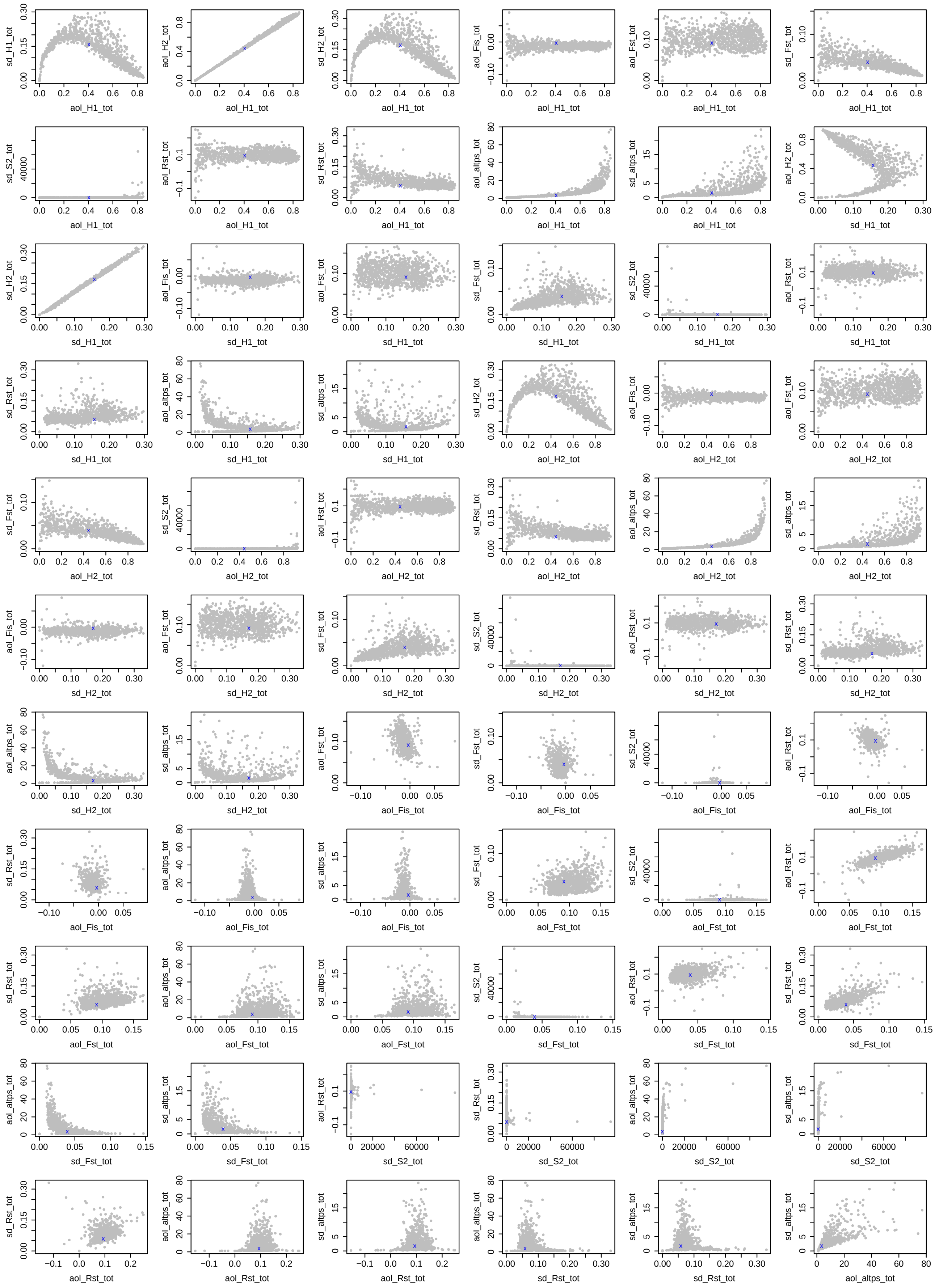

Frequency
p: 0.011
all.loc; ε = 0.001
$\mu_{\theta_{anc}}$
p: 0.61
all.loc; ε = 0.001
$\log_{10}(\sigma_{\theta_{anc}})$
p: 0.859
all.loc; ε = 0.001
$\log_{10}(\omega)$
p: 0.003
all.loc; ε = 0.01
$\mu_{\theta_{anc}}$
p: 0.992
all.loc; ε = 0.01
$\log_{10}(\sigma_{\theta_{anc}})$
p: 0.88
all.loc; ε = 0.01
$\log_{10}(\omega)$
p: 0.003
all.loc; ε = 0.1
$\mu_{\theta_{anc}}$
p: 0.677
all.loc; ε = 0.1
$\log_{10}(\sigma_{\theta_{anc}})$
p: 0.482
all.loc; ε = 0.1
$\log_{10}(\omega)$
Frequency
p: 0.314
pls.loc; ε = 0.001
p: 0.314
p: 0.723
p: 0.023
pls.loc; ε = 0.01
p: 0.232
p: 0.655
p: 0.001
pls.loc; ε = 0.1
p: 0.078
p: 0.273
Frequency
p: 0.006
lgb.loc; ε = 0.001
p: 0.5
p: 0.794
p: 0.001
lgb.loc; ε = 0.01
p: 0.576
p: 0.862
p: 0
lgb.loc; ε = 0.1
p: 0.991
p: 0.423
Frequency
p: 0.062
l1b.loc; ε = 0.001
p: 0.61
p: 0.61
p: 0.012
l1b.loc; ε = 0.01
p: 0.651
p: 0.791
p: 0
l1b.loc; ε = 0.1
p: 0.257
p: 0.694
Frequency
p: 0.314
l2b.loc; ε = 0.001
p: 0.759
p: 0.5
p: 0.084
l2b.loc; ε = 0.01
p: 0.651
p: 0.532
p: 0
l2b.loc; ε = 0.1
p: 0.284
p: 0.579
0.0
0.4
0.8
0
20
40
60
80

Figure S10 ***(facing page)*** Pairwise prior predictive distribution of summary statistics on original scale. Only summary statistics chosen with the $\mathrm{lgb.glob}$ method are shown. Gray points represent $N = 1000$ simulations with parameter values drawn from the prior. The true value from the ibex data set is shown as a blue cross; *aol*, average over loci; *sd*, standard deviation over loci.


sd_H1_tot
aol_H1_tot
aol_H2_tot
aol_H1_tot
sd_H2_tot
aol_H1_tot
aol_Fis_tot
aol_H1_tot
aol_Fst_tot
aol_H1_tot
sd_Fst_tot
aol_H1_tot
sd_S2_tot
aol_H1_tot
aol_Rst_tot
aol_H1_tot
sd_Rst_tot
aol_H1_tot
aol_altps_tot
aol_H1_tot
sd_altps_tot
aol_H1_tot
aol_H2_tot
sd_H1_tot
sd_H2_tot
sd_H1_tot
aol_Fis_tot
sd_H1_tot
aol_Fst_tot
sd_H1_tot
sd_Fst_tot
sd_H1_tot
sd_S2_tot
sd_H1_tot
aol_Rst_tot
sd_H1_tot
sd_Rst_tot
sd_H1_tot
aol_altps_tot
sd_H1_tot
sd_altps_tot
sd_H1_tot
sd_H2_tot
aol_H2_tot
aol_Fis_tot
aol_H2_tot
aol_Fst_tot
aol_H2_tot
sd_Fst_tot
aol_H2_tot
sd_S2_tot
aol_H2_tot
aol_Rst_tot
aol_H2_tot
sd_Rst_tot
aol_H2_tot
aol_altps_tot
aol_H2_tot
sd_altps_tot
aol_H2_tot
aol_Fis_tot
sd_H2_tot
aol_Fst_tot
sd_H2_tot
sd_Fst_tot
sd_H2_tot
sd_S2_tot
sd_H2_tot
aol_Rst_tot
sd_H2_tot
sd_Rst_tot
sd_H2_tot
aol_altps_tot
sd_H2_tot
sd_altps_tot
sd_H2_tot
aol_Fst_tot
aol_Fis_tot
sd_Fst_tot
aol_Fis_tot
sd_S2_tot
aol_Fis_tot
aol_Rst_tot
aol_Fis_tot
sd_Rst_tot
aol_Fis_tot
aol_altps_tot
aol_Fis_tot
sd_altps_tot
aol_Fis_tot
sd_Fst_tot
aol_Fst_tot
sd_S2_tot
aol_Fst_tot
aol_Rst_tot
aol_Fst_tot
sd_Rst_tot
aol_Fst_tot
aol_altps_tot
aol_Fst_tot
sd_altps_tot
aol_Fst_tot
sd_S2_tot
sd_Fst_tot
aol_Rst_tot
sd_Fst_tot
sd_Rst_tot
sd_Fst_tot
aol_altps_tot
sd_Fst_tot
sd_altps_tot
sd_Fst_tot
aol_Rst_tot
sd_S2_tot
sd_Rst_tot
sd_S2_tot
aol_altps_tot
sd_S2_tot
sd_altps_tot
sd_S2_tot
sd_Rst_tot
aol_Rst_tot
aol_altps_tot
aol_Rst_tot
sd_altps_tot
aol_Rst_tot
aol_altps_tot
sd_Rst_tot
sd_altps_tot
sd_Rst_tot
sd_altps_tot
aol_altps_tot

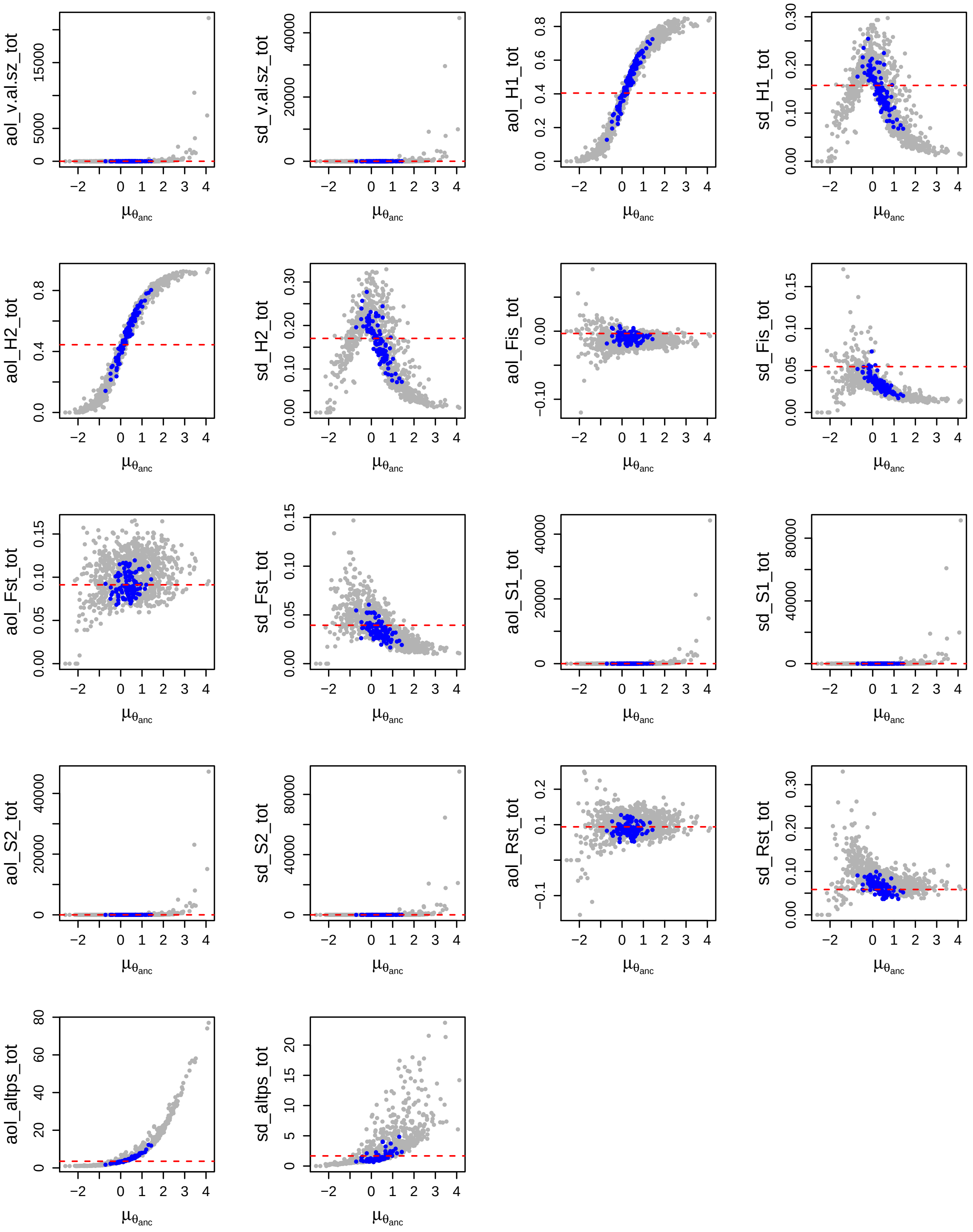


Figure S11 Relation between $\mu_{\theta_{\text{anc}}}$ and the candidate summary statistics. Summary statistics are on the y-axis; *aol*, average over loci; *sd*, standard deviation over loci. Gray points represent $n = 10,000$ simulations, the red dashed line corresponds to the observation for Alpine ibex. Blue points give the $n' = 1000$ simulations closest to the observation, where 'closeness' was defined as described in the main text (*cf.* Figure S2).

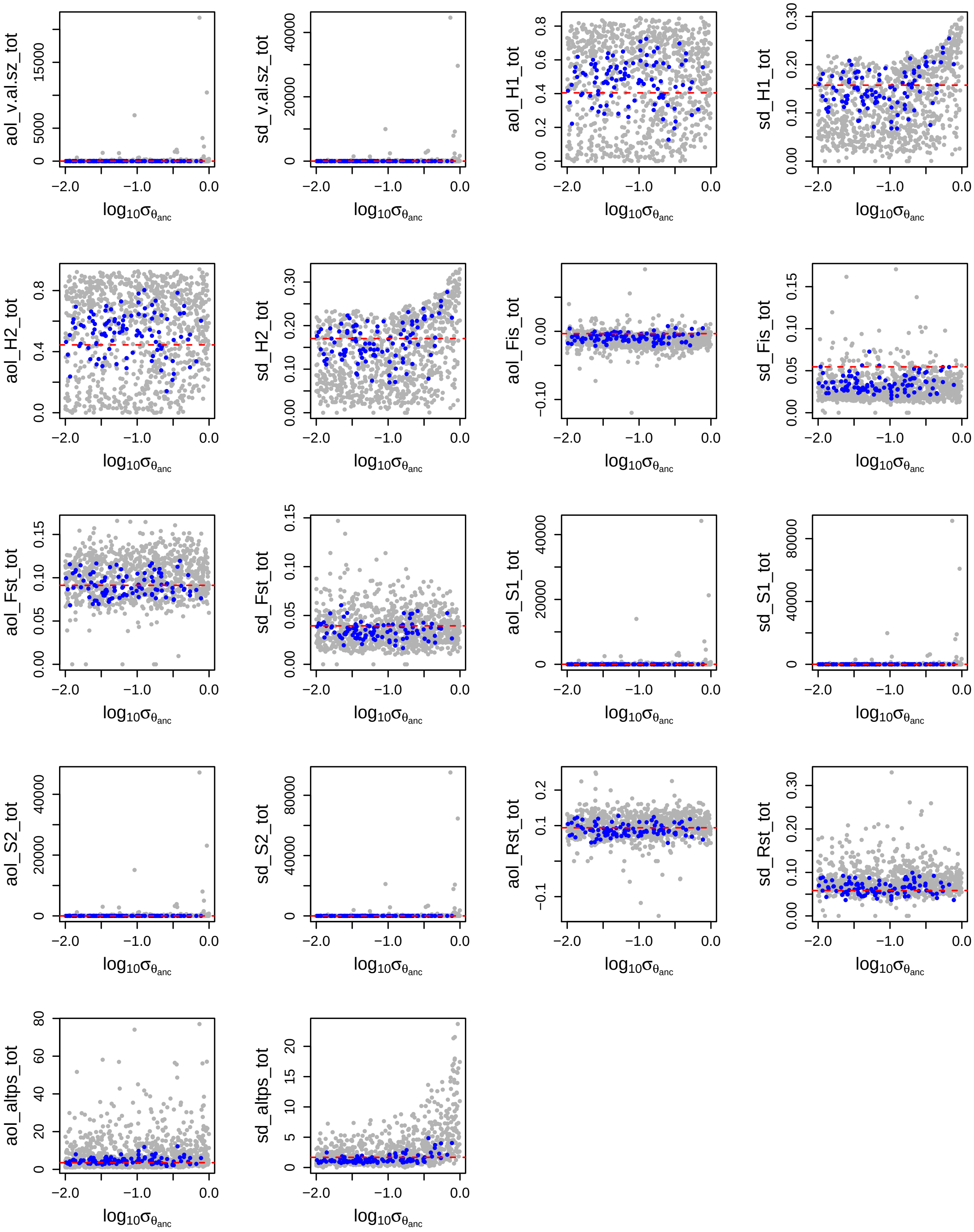


Figure S12 Relation between $\log_{10} \sigma_{\theta_{\mathrm{anc}}}$ and the candidate summary statistics. Details as in Figure S11.

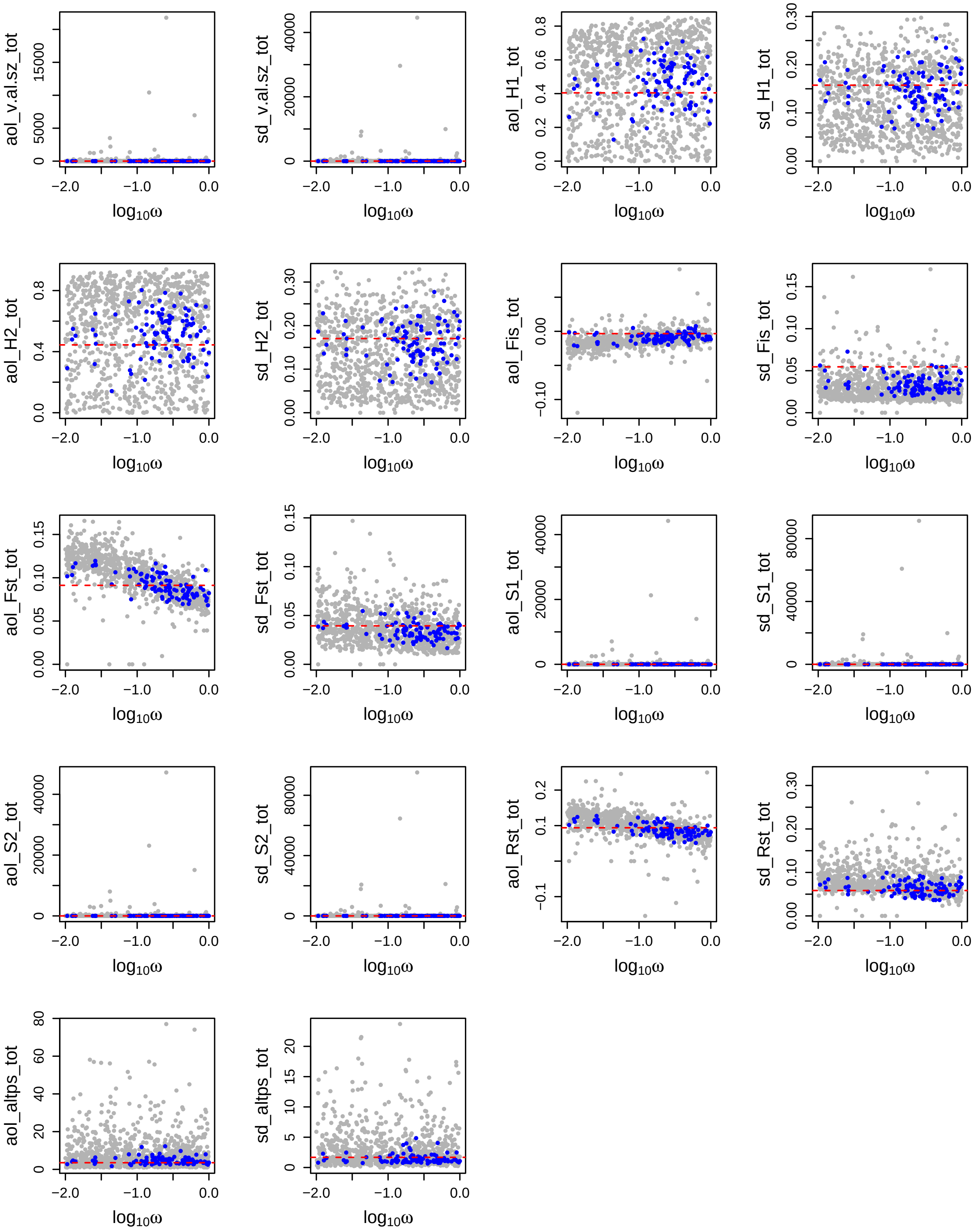


Figure S13 Relation between $\log_{10}\omega$ and the candidate summary statistics. Details as in Figure S11.

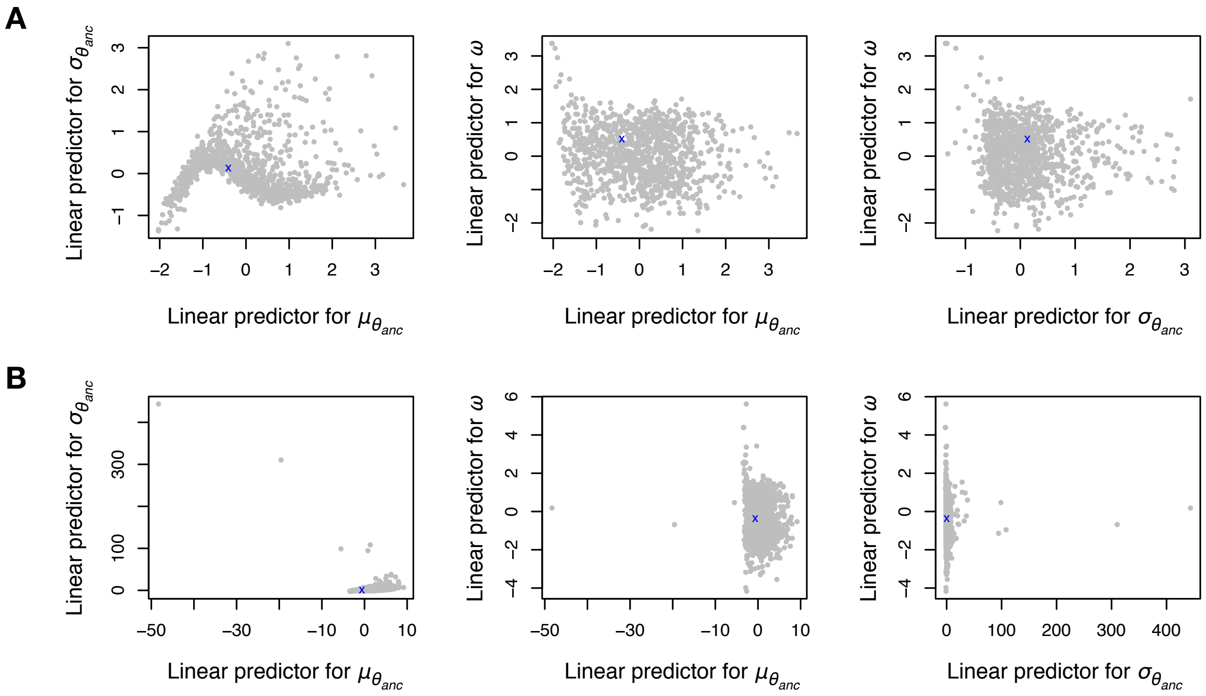


Figure S14 Effect of local choice on scale of summary statistics. Summary statistics were chosen with L_2Boosting as explained in the main text. For each parameter, one linear combination of the original statistics is used as the new summary statistic. These linear combinations are plotted against each other. (A) Global choice of summary statistics. (B) Local choice of summary statistics. Gray points represent $N = 1000$ simulations and the blue cross marks the value observed for Alpine ibex. The local choice of statistics leads to a rescaling compared to the global choice.

SI Supporting Information: Additional Methods

SI.1 Demography and life cycle in simulations

In the following, we give additional details of the demographic model and the ibex-specific settings used in the simulations. All of this is implemented in the program SPoCS (Simulate Populations under Complex Scenarios) written in Java and available on the website http://pub.ist.ac.at/~saeschbacher/phd_e-sources/.

SI.1.1 Life cycle

Alpine ibex is a long-lived, middle-sized ungulate species (Toïgo *et al.* 2002; 2007). We divide the life cycle into years and a year into discrete events, some of which are further described below. We set the maximum age of females and males to 22 and 17 years, respectively (Nievergelt 1966, Toïgo *et al.* 2007). Females and males reach sexual maturity at an age of 3 years (Nievergelt 1966, Stuwe and Grodinsky 1987, Toïgo *et al.* 2002), and the expected age of first reproduction for females and males is 4 and 9 years, respectively (Loison *et al.* 2002, Toïgo *et al.* 2002). In our simulations, females and males stop reproducing when older than 20 and 15 years, respectively.

SI.1.2 Founder/admixture events

A new deme is established by founder individuals taken from previously existing demes. The minimum and maximum age of a founder is 1 and 7 years, respectively, independently of sex. Existing demes may receive further individuals from other demes at later points in time (as specified in Supporting File S3 *transfers*). The range of ages allowed for these admixing individuals is the same as for founders. Founder/admixture events take place at the beginning of the year, before the regulating deaths (see below).

SI.1.3 Reproduction

Females reproduce according to a baseline fertility parameter f. It gives the probability that, for a given year, a particular female will reproduce. If the female reproduces, she mates with a male randomly chosen from the set of males with access to matings in that year (see below). Given a particular female reproduces, it may have one or two offspring. This is controlled by the twin rate parameter $z := \Pr[\text{twins} \mid \text{female reproduces}]$. We set $f = 0.4$ (Nievergelt 1966, Stuwe and Grodinsky 1987) and $z = 0.08$ (Toïgo *et al.* 2002).

Males can get access to matings if they reached the expected age of first reproduction (9 years) and are then counted as potentially reproducing. If, in a deme, no males older than 9 years are available, all males older than the age of sexual maturity (3 years) are considered potentially reproducing. The proportion of these potentially reproducing males that actually get access to matings is defined as ω (see main text). It is one of the parameters to be estimated in this study.

SI.1.4 Deme size control

If the number of offspring required to reach the deme size of the next year cannot be produced by the female baseline fertility f (see above), additional females are allowed to reproduce: Rather than allowing only females to reproduce who reached the expected age of first reproduction (4 years), all females who reached the age of sexual maturity (3 years) may reproduce in this case. If, on the other hand, baseline reproduction results in more individuals than needed to reach the census size of the next year, surplus individuals are removed. These regulating deaths are irrespective of age and sex, and additional to the natural deaths of senescence. In any case, we limit the proportion by which the reproductive need may be overshot per year to 0.2.

SI.1.5 Migration

We simulate migration after the regulating deaths, but before reproduction. Females and males must have reached the age of 3 years before they emigrate (they are then 'potential emigrants'). For a given source deme, the total of individuals to be sent to all connected demes (see main text) are put into an emigrant pool. Emigrants are then randomly distributed to the receiver demes in proportions corresponding to the emigration rates.

SI.2 Explicit forms of minimum expected loss and negative gradient in boosting

The FGD algorithm given in the Appendix of the main text is generic. It is instructive to study the explicit form of expressions in step 1 and 2 of this algorithm for the specific loss functions used here. To this purpose, we follow Friedman *et al.* (2000), Friedman (2001) and Bühlmann and Hothorn (2007).

SI.2.1 Population minimizer of expected loss

We first give explicit forms of the population minimizer (6) for the three loss functions in equations (9), (10) and (12). These are obtained by minimizing the expectation of the joint distribution of $\mathbf{X}$ and Y, $\mathbb{E}_{\mathbf{X},Y}[L(Y,F)]$, where $L(\cdot,\cdot)$ is the generic loss function and $F = F(\mathbf{X})$. In our context, it is enough to take the expectation conditional on $\mathbf{X} = \mathbf{x}$, $\mathbb{E}_Y[L(Y,F) \mid \mathbf{x}]$.

For the L_1-loss in (9), $F^*(\cdot)$ from (6) is obtained as the $F(\cdot)$ that minimizes $\mathbb{E}_Y[|Y-F| \mid \mathbf{x}]$. By the definition of the median, the population minimizer is (Friedman 2001, Bühlmann and Hothorn 2007)

$$F^*(\mathbf{x}) = \mathrm{median}(Y \mid \mathbf{x}). \tag{23}$$

For the L_2-loss in (10), the expected loss is $\mathbb{E}_Y[(Y-F)^2/2 \mid \mathbf{x}]$, and $F^*(\cdot)$ is obtained by setting the derivative with respect to F to zero:

$$\frac{\partial}{\partial F}\,\mathbb{E}_Y\left[\frac{1}{2}\,(Y-F)^2\,\middle|\,\mathbf{x}\right] = \frac{1}{2}\,\frac{\partial\mathbb{E}_Y[Y^2\,|\,\mathbf{x}]}{\partial F} - \frac{\partial\mathbb{E}_Y[Y\,F\,|\,\mathbf{x}]}{\partial F} + \frac{1}{2}\,\frac{\partial\mathbb{E}_Y[F^2\,|\,\mathbf{x}]}{\partial F}$$
$$= 0 - \mathbb{E}_Y[Y\,|\,\mathbf{x}] + F(\mathbf{x}) = 0, \tag{24}$$

from which the familiar result

$$F^*(\mathbf{x}) = \mathbb{E}_Y[Y\,|\,\mathbf{x}] \tag{25}$$

follows (Friedman 2001, Bühlmann and Hothorn 2007).

Friedman *et al.* (2000) show how to derive the population minimizer of the negative binomial log-likelihood in equation (12). For notational convenience, we encode the response by $\tilde{Y} = 2Y-1 \in \{-1,1\}$. The likelihood in (12) can then be written as

$$L(\tilde{Y},F) = \log\left(1+e^{-\tilde{Y}F}\right). \tag{26}$$

In analogy to our previous definition, we set $p(\mathbf{x}) := \Pr[\tilde{Y} = 1\,|\,\mathbf{X} = \mathbf{x}]$, and hence $1-p(\mathbf{x}) := \Pr[\tilde{Y} = -1\,|\,\mathbf{X} = \mathbf{x}]$. Dropping the arguments, we have

$$\mathbb{E}_{\tilde{Y}}[L\,|\,\mathbf{x}] = \mathbb{E}_{\tilde{Y}}\left[\log(1+e^{\tilde{Y}F})\,\middle|\,\mathbf{x}\right]$$
$$= p\log(1+e^{-F}) + (1-p)\log(1+e^{F}). \tag{27}$$

The partial derivative with respect to F is

$$\mathbb{E}_{\tilde{Y}}\left[\log(1+e^{\tilde{Y}F})\,\middle|\,\mathbf{x}\right] = -p\,\frac{e^{-F}}{1+e^{-F}} + (1-p)\,\frac{e^{F}}{1+e^{F}}. \tag{28}$$

Setting to zero and solving for F, we obtain the population minimizer

$$F^*(\mathbf{x}) = \log\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right]. \tag{29}$$

Notice that Friedman *et al.* (2000) and Bühlmann and Hothorn (2007) use a slightly different parameterization, namely setting F equal to *one half* of the logit-transform, such as to have the population minimizer equal to the one for the exponential loss criterion. The population minimizers in (23), (25) and (29) imply that the initial function estimates in step 1 of the FGD algorithm (Appendix) must be set to $F^*(\cdot) \equiv \mathrm{median}(Y)$ for the L_1-loss, to $F^*(\cdot) \equiv \bar{Y}$ for the L_2-loss, and to $F^*(\cdot) \equiv \log[\hat{p}/(1-\hat{p})]$ for the negative binomial log-likelihood loss.

SI.2.2 Negative gradient

To calculate the negative gradient vector $(U_1,\ldots,U_n)$ in step 2 of the FGD algorithm (Appendix), we need the partial derivative of the loss function with respect to the target function F. Any element U_i is obtained as this partial derivative evaluated at the previous function estimate $\hat{F}^{[m-1]}(\mathbf{x}_i)$. Formally,

$$U_i = -\frac{\partial}{\partial F} L(Y_i, F)\bigg|_{F=\hat{F}^{[m-1]}(\mathbf{X}_i)}. \tag{30}$$

For the L_1-loss in (9), we have

$$-\frac{\partial}{\partial F}\Big[|Y_i - F|\Big] = \frac{Y_i - F}{|Y_i - F|} = \mathrm{sgn}(Y_i - F), \tag{31}$$

which implies the negative gradient component

$$U_i = \mathrm{sgn}\Big[Y_i - \hat{F}^{[m-1]}(\mathbf{X}_i)\Big] \tag{32}$$

in step 2 of the FGD algorithm (*cf.* Friedman 2001).

For the L_2-loss in (10),

$$-\frac{\partial}{\partial F}\left[\frac{1}{2}(Y_i - F)^2\right] = Y_i - F, \tag{33}$$

which amounts to

$$U_i = Y_i - \hat{F}^{[m-1]}(\mathbf{X}_i) \tag{34}$$

in step 2 of the FGD algorithm (*cf.* Friedman 2001, Bühlmann and Hothorn 2007).

Last, for the negative binomial log-likelihood we again use $\tilde{Y} = 2Y - 1 \in \{-1, 1\}$ and find

$$-\frac{\partial}{\partial F} L(\tilde{Y}_i, F) = -\frac{\partial}{\partial F}\log\left(1 + e^{-\tilde{Y}_i F}\right) = \frac{\tilde{Y}_i\, e^{-\tilde{Y}_i F}}{1 + e^{-\tilde{Y}_i F}}. \tag{35}$$

This leads to the negative gradient component

$$U_i = \frac{\tilde{Y}_i\, e^{-\tilde{Y}_i \hat{F}^{[m-1]}(\mathbf{X}_i)}}{1 + e^{-\tilde{Y}_i \hat{F}^{[m-1]}(\mathbf{X}_i)}} \tag{36}$$

in step 2 of the FGD algorithm.

SI Supporting Information: URLs to Supporting Files

File S2
Census population sizes of Alpine ibex demes in the Swiss Alps

File S2 is available for download as a PDF file at http://pub.ist.ac.at/~saeschbacher/phd_e-sources/[to be found in subsection 'Chapter 3'].

File S3
Numbers of Alpine ibex transferred between demes by humans

File S3 is available for download as a PDF file at http://pub.ist.ac.at/~saeschbacher/phd_e-sources/[to be found in subsection 'Chapter 3'].

Literature Cited

Bühlmann, P. and T. Hothorn, 2007 Boosting algorithms: Regularization, prediction and model fitting. Stat. Sci. **22**: 477–505.

Friedman, J., T. Hastie, and R. Tibshirani, 2000 Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting. Ann. Stat. **28**: 337–374.

Friedman, J. H., 2001 Greedy function approximation: A gradient boosting machine. Ann. Stat. **29**: 1189–1232.

Loison, A., C. Toïgo, J. Appolinaire, and J. Michallet, 2002 Demographic processes in colonizing populations of isard (*Rupicapra pyrenaica*) and ibex (*Capra ibex*). J. Zool. **256**: 199–205.

Nievergelt, B., 1966 *Der Alpensteinbock (*Capra ibex L.*) in seinem Lebensraum. Ein ökologischer Vergleich*. Mammalia depicta, Verlag Paul Parey, Hamburg, Berlin.

Stuwe, M. and C. Grodinsky, 1987 Reproductive biology of captive Alpine ibex (*Capra i. ibex*). Zoo Biol. **6**: 331–339.

Toïgo, C., J. M. Gaillard, M. Festa-Bianchet, E. Largo, J. Michallet, and D. Maillard, 2007 Sex- and age-specific survival of the highly dimorphic Alpine ibex: evidence for a conservative life-history tactic. J. Anim. Ecol. **76**: 679–686.

Toïgo, C., J. M. Gaillard, D. Gauthier, I. Girard, J. P. Martinot, and J. Michallet, 2002 Female reproductive success and costs in an alpine capital breeder under contrasting environments. Écoscience **9**: 427–433.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182**: 1207–1218.