Selection of Pairings Reaching Evenly Across the Data (SPREAD): A simple algorithm to design maximally informative fully crossed mating experiments

Kolea Zimmerman[†], Daniel Levitis[§], Ethan Addicott[†], and Anne Pringle[‡]

† Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, 02138

§ Max-Planck Odense Center on the Biodemography of Aging, Institute of Biology, University of Southern Denmark, DK-5230 Odense M, Denmark

‡ Harvard Forest, Harvard University, Petersham MA, 01366

Corresponding Author: Kolea Zimmerman

ABSTRACT

We present a novel algorithm for the design of crossing experiments. The algorithm identifies a set of individuals (a "crossing-set") from a larger pool of potential crossing-sets by maximizing the diversity of traits of interest, for example, maximizing the range of genetic and geographic distances between individuals included in the crossing-set. To calculate diversity, we use the mean nearest neighbor distance of crosses plotted in trait space. We implement our algorithm on a real dataset of *Neurospora crassa* strains, using the genetic and geographic distances between potential crosses as a two-dimensional trait space. In simulated mating experiments, crossing-sets selected by our algorithm provide better estimates of underlying parameter values than randomly chosen crossing-sets.

INTRODUCTION

Researchers planning mating experiments are faced with a critical design choice – deciding how many pairs and which pairs of individuals to mate. The number of crosses in a mating experiment can influence statistical estimates of genetic effects and combining abilities (Jui and Lefkovitch, 1992). The selection of pairs to use in a mating experiment also affects the outcome of the experiment. For example, if the goal of a mating experiment is to understand the genetic basis of a trait, as in quantitative trait locus (QTL) analysis, then parents should carefully be chosen to maximize genetic diversity and increase the likelihood of detecting QTLs (Crepieux *et al.*, 2004). The increasing accessibility of population genetic and genomic datasets offer genetic data on more individuals than can reasonably be used in most experiments (Cushman, 2014). This poses a methodological problem: how to choose a subsample of mating pairs that best reflects the range of cross characteristics (in two or more dimensions of genetic, geographic, or ecological space) of the complete set of all available pairs.

One solution is to select a subsample that recapitulates characteristics of the larger set and preserves underlying relationships between the variables used to define a trait space. The representative subsample might mimic the broad distribution of crosses in the larger set, in other words, attempt to maintain the

shape, clumps, etc. of the larger set. The best method to generate a truly representative subsample is not obvious.

Samples chosen by eye or randomly may truncate the trait space, perhaps because a subsample omits outliers or is disproportionately drawn from the dense center of a distribution. Omissions in sampling may hinder a complete understanding of how response variables, for example reproduction, vary across the trait space of all possible crosses. Furthermore, predicting response variables outside the range of independent variables used in an experiment involves extreme value methods, which can increase the error associated with predictions, unless limiting assumptions are made (Pauli and Coles, 2001).

Often, fully crossed experiments are desirable (Griffing, 1956); (Verhoeven *et al.*, 2005). But directed subsampling of potential crosses with the aim of maximizing e.g. genetic diversity may result in a set of crosses that are not fully crossed, i.e., some females included in the mating experiment may not be mated to all included males. A method to subsample and achieve a mating design that is both broad (in the sense described above) and fully crossed is required.

Algorithms for maximizing combinatorial diversity have been extensively developed in the context of generating diverse molecular libraries

for drug screening (Martin and Critchlow, 1999). In these algorithms the metric of diversity is based on "redundancy" and "coverage" (Martin and Critchlow, 1999). Redundancy is the overlapping or clumping of points in space, while coverage is the spread of points across the space. An ideal diversity metric would minimize redundancy while maximizing coverage. The algorithms used in chemical combinatorial analysis focus on maximizing the diversity of a subset of molecules from a larger set by step-wise analysis of differences between additional compounds added to a set (Holliday *et al.*, 1995). These algorithms cannot be directly applied to our problem because they do not require selection of fully crossed sets. However, we use their definitions of ideal set diversity to derive our own measure of diversity that can be applied to fully crossed sets.

Calculating the mean of the nearest neighbor distances (NND) of points representing a full factorial set of crosses plotted based on their underlying parameters (e.g. genetic, geographic or ecological distance) will give a measure of the evenness or "non-redundancy" of the points. The mean NND is often used to determine if a particular set of plotted points is randomly distributed or not (Clark and Evans, 1954). A set of plotted points that are clumped will result in a smaller value of the mean NND than a sample with the same number of more evenly and broadly distributed points. The maximum mean NND

(MMNND) will occur when points are spread as evenly as possible and the "coverage" of space is maximal (Wang and Cumming, 2011). Thus, identifying the set of crosses with the MMNND within an array of many potential sets of crosses ("crossing-sets") will return a crossing-set that is both broad and even with respect to underlying trait values.

In this note, we introduce a simple algorithmic sampling method for choosing crossing-sets; we name the algorithm SPREAD (Selection of Pairings Reaching Evenly Across the Data). SPREAD is based on finding the single crossing-set with the MMNND among many different potential crossing-sets plotted on two-dimensional trait space. We use our algorithm to select a crossing-set from a genotyped collection of geographically widespread wild strains of the filamentous fungus, *Neurospora crassa*. Strains of this fungus have one of two mating types, denoted mat-A or mat-a. The two parents in a cross must have different mating types to mate. Recently, 24 strains of each mating type were genotyped using RNAseq (Ellison *et al.*, 2011). The genotyped strains were collected from diverse locations, allowing us to assign both genetic (the number of different SNPs) and geographic (the distance between collection sites) distance values to each of the 576 potential crosses. Using this dataset as our example, we implemented the SPREAD algorithm and tested the effectiveness of the SPREAD algorithm when the true MMNND is

not easily calculable. We then used a simulated dataset with known underlying parameter values that relate genetic and geographic distances to reproductive output to compare SPREAD to simple random sampling (SRS).

METHODS

Description of the SPREAD Algorithm: Define $\mathbb{X}$ and $\mathbb{Y}$ as the set of available strains or individuals of each mating type or sex ('type') *x* and *y* respectively, and $s_x \times s_y$ as the feasible number of crosses that can be completed in an experiment. The variables s_x and s_y are the number of strains selected for the experiment and are less than $|\mathbb{X}|$ and $|\mathbb{Y}|$ respectively. Draw a large number, *h*, of random samples containing s_x and s_y strains of each type from all possible sets of strains $\begin{pmatrix} \mathbb{X} \\ s_x \end{pmatrix}$ and $\begin{pmatrix} \mathbb{Y} \\ s_y \end{pmatrix}$. For each of the *h* samples, plot crosses based on values associated with the crosses (for example number of differing SNPs vs. geographic distance), and then calculate the mean of the NNDs of all plotted crosses. Generate a list of *h* mean NNDs. Finally, use the maximum value from the list because it corresponds to the set of $s_x \times s_y$ strains that most broadly and evenly represents the parameter of interest. A formal mathematical description of this algorithm is available online at http://dx.doi.org/10.6084/m9.figshare.1180170.

A worked example using SPREAD: We used a previously published population genomics dataset consisting of single nucleotide polymorphisms (SNPs) from transcriptomes of geographically diverse wild isolates of the fungus *N. crassa* to test our method (Ellison *et al.*, 2011). We started with the set of all pairwise combinations of strains and then filtered to include only mating type compatible pairs. We calculated genetic distances between compatible pairs by counting the number of different SNPs between each pair and calculated geographic distances using the great-circle distance between strain locales. The genetic and geographic distance values for each pair were used to map all the crosses on genetic and geographic distance axes. This is the "original distribution" of crosses. We randomly sampled h = 1000 lists of s strains of each mating type from the set of all $\begin{pmatrix} \mathbb{A} \\ s_a \end{pmatrix}$ and $\begin{pmatrix} \mathbf{a} \\ s_a \end{pmatrix}$ strains without replacement, where $\mathbb{A}$ and $\mathbf{a}$ and are the sets of strains available for each mating type; in this case $|\mathbb{A}| = |\mathbf{a}| = 24$ and $s_A = s_a = 12$. We computed all possible pairwise mating combinations for each of the random samples of $s_A = s_a = 12$ strain lists, resulting in 1000 crossing-sets each containing 144 crosses. We then plotted each crossing-set on geographic vs. genetic distance space and computed the mean nearest neighbor distances using Euclidean

distance calculations. The crossing-set with the MMNND of all 1000 crossing-sets was selected.

We implemented our algorithm and additional analyses in the R programming language (R Core Team, 2014). R code for the implementation of the SPREAD algorithm on crossing-sets with two traits is available online at http://dx.doi.org/10.6084/m9.figshare.1180165. The following R packages were used in this analysis: plyr (Wickham, 2011), reshape2 (Wickham, 2007), ggplot2 (Wickham, 2009), spatstat (Baddeley and Turner, 2005), Rmpi (Yu, 2002), doMPI (Weston, 2013), doRNG (Gaujoux, 2014), foreach (Weston and Analytics, 2014), and glmmADMB (Fournier *et al.*, 2012; Skaug *et al.*, 2013).

Evaluating SPREAD's approximation of the true MMNND: The true MMNND can only be determined if all possible crossing-sets for a given s_A and s_a are evaluated. Therefore, calculating the true MMNND may not be possible, even with high performance computing resources. For example, if M = 300 and F = 300 and a crossing set is desired with 20 individuals of each type, then the total number of possible crossing sets would be $\binom{300}{20}^2 = 5.6 \times 10^{61}$. Using a random sample of all available crossing-sets to estimate the MMNND would be more practical, especially if the estimated MMNND approximates the true MMNND.

We implemented the algorithm as described above for the *N. crassa* dataset, except we varied crossing-set size by implementing SPREAD for $s_A = s_a = 3, 4, 5, \ldots, 21$. To simplify the process, we used crossing-sets where $s_A = s_a$, but this is not a requirement of the SPREAD algorithm. We used two different *h* values, 100 and 1000, to compare the effects of *h* size on the MMNND values returned from SPREAD. We repeated this process 1000 times to obtain bootstrapped distributions of MMNND values for the different crossing-set sizes and *h* values.

Comparing model fits of SPREAD and SRS generated crossing-sets: Using SPREAD to design fully crossed mating experiments may be more effective than selecting crossing-sets at random because broad and even sampling will provide greater power to understand how dependent variables vary based on cross characteristics (e.g. how reproductive success depends on the genetic or geographic distances between parents). To evaluate this hypothesis empirically, we created a simulated dataset of cross outcomes, i.e. reproduction, and modeled relationships between reproduction and the characteristics of crosses in crossing-sets generated from SPREAD *versus* those generated by simple random sampling. Simulated experimental data take the form of total ascospore counts.

First, we generated simulated data for all possible crosses of the entire crossing-set of 24 mat-A $\times$ 24 mat-a strains, using a generalized linear model (GLM) fitted to unpublished empirical data. The model is described as follows:

$$Y_i = \beta_0 + X_i\beta_1 + Z_i\beta_2 + X_i^2\beta_3 + Z_i^2\beta_4 + X_iZ_i\beta_5 + \varepsilon_i$$

where Y is total ascospore count, X is genetic distance between a cross, and Z is geographic distance between a cross. This model was evaluated using the glmmADMB package (Fournier *et al.*, 2012; Skaug *et al.*, 2013) as *Total Ascospore Count* = *Genetic Distance* + *Geographic Distance* + (*Genetic Distance*)^2 + (*Geographic Distance*)^2 + *Genetic Distance : Geographic Distance*. The response variable of *Total Ascospore Count* was modeled with a negative binomial distribution using a log link function.

Second, we simulated four experimental replicates for each possible cross by drawing from a negative binomial distribution with a mean derived from the predicted experimental values and the negative binomial dispersion parameter derived from the empirical data model. “True” parameter values were determined by fitting the complete simulated data set of all crosses to the model described above.

Using this complete set of simulated experimental data we calculated model fits for 1000 different crossing-sets generated with SPREAD and, then, SRS. The algorithm parameter values for the SPREAD generated crossing-sets

were $s_A = s_a = 12$ and $h = 1000$. We chose $s_A = s_a = 12$ to test the edge case of a maximally complex sample space (the largest number of possible crossing-set permutations occurs when $s_A = s_a = 12$). Crossing-sets chosen by SRS were of the same size. Model fits were computed for each crossing-set using the model described above. Parameter values and standard errors of the parameter values were recorded for each of the 1000 SPREAD or SRS generated crossing-sets.

RESULTS

The worked example: We used SPREAD on the *N. crassa* dataset described above to select a crossing-set with 12 mat-A and 12 mat-a strains. A graphical assessment of the chosen crossing-set plotted on geographic and genetic distance axes shows that our method produces a crossing-set that broadly and evenly represents all potential crosses (Figure 1).

Implementing SPREAD without knowing the true MMNND: Plotting estimates of MMNND for both *h* values (the number of randomly sampled crossing-sets from which the set with MMNND is chosen) shows that the variance of the MMNND values is larger for $h = 100$ than $h = 1000$ (Figure 2). However, the variance in the MMNND values rapidly decreases as the size of the crossing-sets increases. The max of the 1000 MMNND values for $h = 100$ is often less than for $h = 1000$ for smaller crossing-set sizes. The difference

in the max MMNND becomes negligible for crossing-sets with more than 100 crosses.

Comparing SPREAD to SRS: The distributions of model fits for crossing-sets generated by SPREAD or SRS and the true parameter values derived from the entire simulated dataset are shown in Figure 3. The peaks of the distributions of parameter values for both SPREAD and SRS selected crossing-sets do not perfectly align with the true parameter values. However, the parameter values from models fitted using crossing-sets generated by SPREAD are distributed more closely around the true parameter values (Figure 3a and Table 1). The standard errors of parameters from model fits of SPREAD generated crossing-sets are smaller and less variable compared to the standard errors of parameter values from model fits of SRS generated crossing-sets (Figure 3b and Table 2).

DISCUSSION

SPREAD is an easily implemented algorithm designed to identify maximally informative, full factorial crossing-sets for use in mating experiments. SPREAD takes information about all of the potential crosses from a genotyped or otherwise characterized population and maximizes the diversity inherent in a crossing-set, for example, the genetic and geographic distances among crosses. SPREAD requires two input parameters chosen by the user: the

dimensions, $s_x \times s_y$, of the desired crossing-set and the number of randomly generated crossing-sets, *h*, used to calculate MMNND and find the ideal crossing-set. SPREAD was designed for two dimensional trait data. If potential crosses are characterized by more than two target traits, and the traits are not completely independent, principal components analysis (PCA) can be used before implementing SPREAD to determine which two traits explain most of the trait variance (King and Jackson, 1999).

In our worked example, we successfully used SPREAD to select a crossing-set of 12 mat-A $\times$ 12 mat-a *N. crassa* strains from a larger set. When these crosses are plotted on genetic vs. geographic distance space, it is evident that the selected set fulfills the desired criteria of evenly and completely covering the range of the larger set (Figure 1). Using the MMNND as the diversity metric favors crosses that are at the extremes of the trait-space. The inclusion of crosses with extreme trait distances in an original population should be carefully considered because these crosses will often be selected by SPREAD.

Calculating the true MMNND by computing MNND values for every possible subset of a sampled population may not be possible. Instead, our algorithm generates *h* subsets and chooses the set with the MMNND from those *h* subsets. Our results show that the MMNND values returned by SPREAD

using an *h* value of 100 or 1000 are minimally variable at all but the smallest crossing-set sizes (Figure 2), and using these *h*-values should be sufficient for most experiments with modest population sizes.

In experiments where the sampled population is very large, the variance in MMNND values may not rapidly decrease with increasing crossing-set size. In these cases greater *h* values should be used. Alternatively, SPREAD could be modified to use a simulated annealing approach to search the space of potential crossing-sets for a crossing-set that converges on a peak MNND value. One example of a simulated annealing algorithm that could be adapted for this purpose is SAGE (Simulated Annealing Guided Evaluation), developed to design combinatorial drug libraries (Zheng *et al.*, 1999).

We hypothesized that maximizing the diversity inherent in a crossing-set would increase the predictive ability of models relating outcomes, for example reproduction, to characteristics of crosses, for example the genetic distances between crosses. When we compared model fits from crossing-sets generated by SPREAD to model fits from crossing-sets generated by SRS, we found that the model parameter values from SPREAD generated crossing-sets were closer to the true model parameter values with smaller errors (Figure 3). Although the parameter values from SPREAD generated crossing-sets were closer to the true parameter values they did not precisely match the true

parameter values. This is probably because the true parameter values are from a model calculated using the entire set of 24 mat-A × 24 mat-a crosses while the tested parameter values are from crossing-sets of 12 mat-A × 12 mat-a strains. The smaller sample size used to fit the model decreases both the precision and accuracy of estimated parameter values. Generalized linear models have been shown to be especially sensitive to sample size, compared to other methods (Wisz *et al.*, 2008).

SPREAD increases the value of fully crossed mating designs by enabling exploration and prediction across the full space of cross characteristics provided by available breeding stock. Simulations based on crossing-sets generated from the SPREAD algorithm *versus* SRS prove our algorithm generates more accurate parameter estimates, enabling better predictions of relationships between cross characteristics (e.g. the genetic and geographic distances between parents) and the success of a cross. SPREAD is not computationally intense and is easy to implement, making it a valuable tool for researchers designing crossing experiments.

ACKNOWLEDGEMENTS

We thank Kareem Carr and Steven Worthington for assistance in designing statistical analyses. We are grateful to members of the Pringle lab for advice and discussion. Our work is supported by the National Science Foundation Graduate Research Fellowship under Grant Nos. (DGE0644491 and DGE1144152) awarded to K.Z. and by other National Science Foundation grants awarded to the Pringle Laboratory. This work was also supported by funds from the Max Planck Institute for Demographic Research to K.Z., D.L., and A.P. Some computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. The work of D.L. was supported by the Max-Planck Odense Center, a collaboration of the Max Planck Society and the University of Southern Denmark.

FIGURES

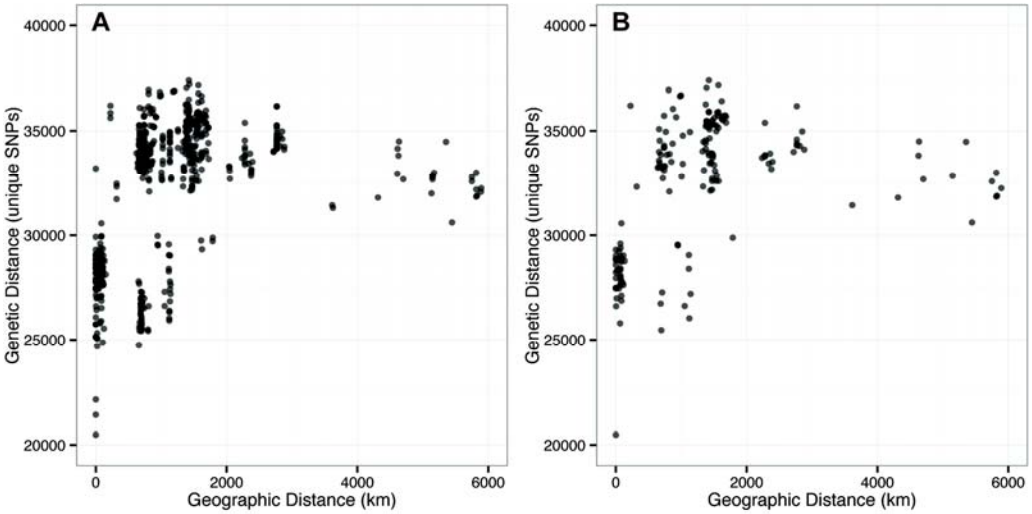

Figure 1 Implementation of SPREAD. Panel **A** shows all possible crosses between 24 mat-A and 24 mat-a strains; panel **B** shows the $s_A = s_a = 12$ crossing-set of *N. crassa* strains returned from SPREAD (with $h = 1000$), where s_A is the number of mat-A strains and s_a is the number of mat-a strains. Crosses are plotted as semitransparent dots and darker colors mark overlapping crosses.

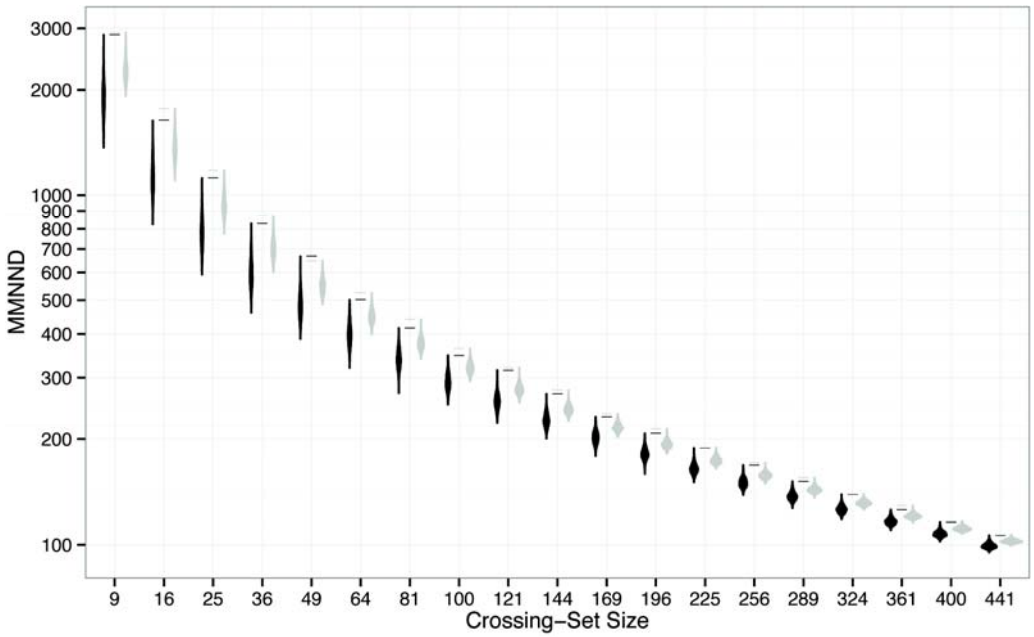

Figure 2 Effect of crossing-set size and *h* value (black: *h* = 100, gray: *h* = 1000) on distributions of MMNND values. With larger *h*-values, estimates of MMNND values increase. As crossing-set size increases MMNND values decrease because with more crosses the average distance between crosses decreases. Violin plots show the entire distributions of MMNND values plotted on a log scale. Dashes mark the maximum value of the distributions.

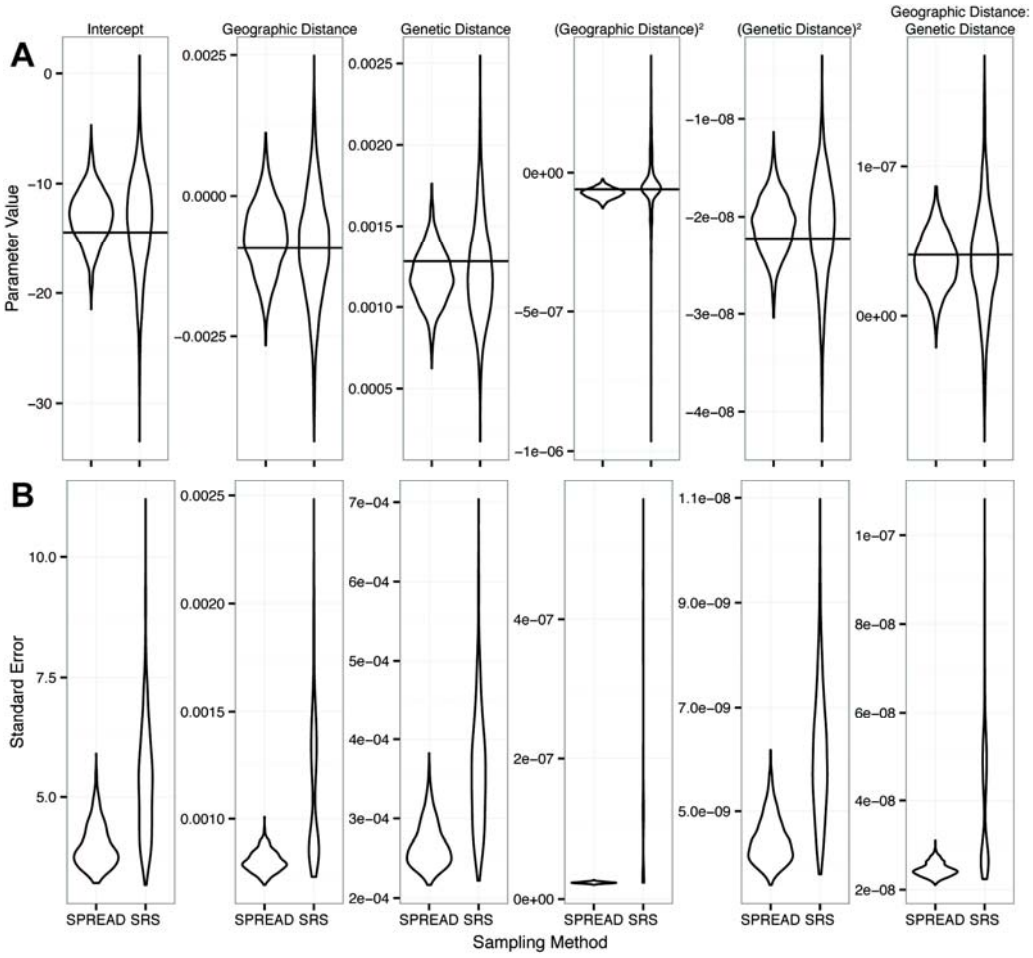

Figure 3 Comparisons of estimated parameter values and standard errors for 1000 crossing-sets generated either with SPREAD or SRS. The SPREAD parameters used are $s_A = s_a = 12$ and $h = 1000$. Panel **A**: Violin plots showing the distribution of parameter values for all terms in the model. Horizontal lines indicate the "true" parameter values of the entire simulated $s_A = s_a = 24$ dataset. Panel **B**: Violin plots showing the distributions of standard errors of all terms in the model.

Table 1 Proportion of crossing-sets with parameter values within the range defined by the true parameter value ± 3 standard errors of true parameter values. "Dist." is the abbreviation for "Distance."

	Proportion of crossing-sets	
Parameter	**SPREAD**	**SRS**
Intercept	0.1527	0.1195
Genetic Distance	0.1513	0.1212
Geographic Distance	0.6422	0.6158
(Genetic Distance)^2	0.1530	0.1253
(Geographic Distance)^2	0.1267	0.0897
Genetic Dist.:Geographic Dist.	0.1475	0.1182

Table 2 Means and variances of standard errors from 1000 model fits using SPREAD or SRS generated crossing-sets. "Dist." is the abbreviation for "Distance."

	Mean of Std. Errors		**Variance of Std. Errors**	
Parameter	**SPREAD**	**SRS**	**SPREAD**	**SRS**
Intercept	3.99E+00	5.33E+00	2.05E-01	1.43E+00
Genetic Distance	2.68E-04	3.61E-04	8.02E-10	5.84E-09
Geographic Distance	8.05E-04	1.14E-03	2.33E-09	1.04E-07
(Genetic Distance)^2	4.45E-09	6.09E-09	1.97E-19	1.55E-18
(Geographic Distance)^2	2.34E-08	9.23E-08	1.44E-18	8.71E-15
Genetic Dist.:Geographic Dist.	2.46E-08	3.92E-08	2.43E-18	2.21E-16

REFERENCES

Baddeley A, Turner R (2005). Spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* **12**: 1–42.

Clark PJ, Evans FC (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology* **35**: 445–453.

Crepieux S, Lebreton C, Servin B, Charmet G (2004). Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. *Genetics* **168**: 1737–1749.

Cushman SA (2014). Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Frontiers in Genetics* **5**: 1–5.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, *et al.* (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc Natl Acad Sci USA* **108**: 2831–2836.

Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, *et al.* (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27**: 233–249.

Gaujoux R (2014). doRNG: Generic Reproducible Parallel Backend for foreach Loops.

Griffing B (1956). Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. *Australian Journal of Biological Sciences* **9**: 463–493.

Holliday JD, Ranade SS, Willett P (1995). A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quantitative Structure-Activity Relationships* **14**: 501–506.

Jui PY, Lefkovitch LP (1992). Selecting the size of a diallel cross experiment. *Theor Appl Genet* **85**: 21–25.

King JR, Jackson DA (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics* **10**: 67–77.

Martin EJ, Critchlow RE (1999). Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J Comb Chem* **1**: 32–45.

Pauli F, Coles S (2001). Penalized likelihood inference in extreme value analyses. *Journal of Applied Statistics* **28**: 547–560.

Skaug H, Fournier D, Nielsen A, Magnusson A, Bolker B (2013). Generalized Linear Mixed Models using AD Model Builder.

Verhoeven KJF, Jannink J-L, McIntyre LM (2005). Using mating designs to uncover QTL and the genetic architecture of complex traits. *Heredity* **96**: 139–149.

Wang X, Cumming SG (2011). Measuring landscape configuration with normalized metrics. *Landscape Ecol* **26**: 723–736.

Weston S (2013). doMPI: Foreach parallel adaptor for the Rmpi package. *R package version 02*.

Weston S, Analytics R (2014). foreach: Foreach looping construct for R.

Wickham H (2007). Reshaping data with the reshape package. *Journal of Statistical Software* **21**: 1–20.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer: New York.

Wickham H (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* **40**: 1–29.

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, *et al.* (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions* **14**: 763–773.

Yu H (2002). Rmpi: Parallel Statistical Computing in R. *R News* **2**: 10–14.

Zheng W, Cho SJ, Waller CL, Tropsha A (1999). Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J Chem Inf Model* **39**: 738–746.