Maximizing the mean nearest neighbor distance of a trait to choose among potential crosses and design a fully crossed mating experiment

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

Short running title: Algorithmic design of crossing schemes

Keywords: mating experiment, cross design, nearest neighbor distance, full factorial cross, subsampling

Corresponding Author: Kolea Zimmerman

Address: 16 Divinity Ave, Biolabs 2112, Cambridge MA 02138

Phone: 617-496-5540

Email: kzimmerman@fas.harvard.edu

ABSTRACT

We present a novel algorithm for the design of crossing experiments. A set of individuals (a "crossing-set") is chosen from a larger set of potential crossing-sets by maximizing the distribution of a trait of interest. In simulated mating experiments, identified crossing-sets provide better estimates of underlying parameter values than randomly chosen crossing-sets.

INTRODUCTION

Designing a mating experiment requires deciding how many and which pairs of individuals to mate. If the goal of a mating experiment is understanding the genetic basis of a trait, as in QTL analysis, then parents should be carefully chosen to maximize genetic diversity and increase the likelihood of detecting QTLs (CREPIEUX *et al.* 2004). In addition to the genetic composition of parents, the number of crosses in a mating experiment will also influence the degree of confidence in parameter estimates (JUI and LEFKOVITCH 1992). 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). The wealth of potential crosses poses a critical methodological question: 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 approach to the problem is to select a sample of crosses and attempt to recapitulate characteristics of the larger set to preserve underlying relationships between the variables used to define trait space. The representative sample might attempt to mimic

the broad distribution of points in the larger set, in other words, attempt to maintain the shape, clumps, etc. of the larger set. However, it is not clear how one might choose a truly representative sample. Samples chosen by eye or randomly may truncate the trait space, perhaps because a subsample omits outliers or is limited to the dense center of a distribution. Omissions in sampling may hinder a complete understanding of how response variables, for example reproductive output, vary across the trait space of all possible crosses. 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). When the aim is to understand how some outcome varies depending on the individuals involved in a cross—for example their genotype and the genetic distance involved in a particular cross—then choosing a set of crosses spread as widely and as evenly across the parameter of interest as possible will be valuable.

Often, fully crossed experiments are desirable because a factorially complete set is required to analyze general combining ability and parental effects (GRIFFING 1956); and specific genetic effects (VERHOEVEN *et al.* 2005). To illustrate the complexity of choosing individuals to maximize representation of e.g. genetic distance among individuals, consider the following example: a set of genotyped individuals includes $F = 40$ females and $M = 30$ males, and the total number of possible crosses in a full-factorial design is $F \times M = 1200$. Directed subsampling of potential crosses with the aim of broad coverage of the genetic distance distribution 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.

Calculating the mean of the nearest neighbor distances (NND) of points representing all potential crosses plotted based on their underlying parameters (e.g. genetic, geographic or ecological distance) will give a measure of the evenness 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, and the maximum mean NND (MMNND) will occur when points are spread as evenly as possible (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 new algorithmic sampling method we name SPREAD (Selection of Pairings Reaching Evenly Across the Data). SPREAD is based on finding the crossing-set with the MMNND among many different potential crossing-sets plotted on n-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.* The fungus is an obligate outcrosser with two mating types, denoted mat-A and 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 a genetic distance value (number of different SNPs) and a geographic distance

value (distance between collection sites) 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 random sampling.

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 $\binom{\mathbb{X}}{s_x}$ and $\binom{\mathbb{Y}}{s_y}$. 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 mathematical formalization of this algorithm is available at http://dx.doi.org/10.6084/m9.figshare.1180170 (ZIMMERMAN *et al.*).

RESULTS

A 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 qualitative assessment of the chosen crossing-set plotted on geographic and genetic distance axes suggests our

method produces a subsample of potential crosses that broadly and evenly represents the crossing-set of all potential crosses (Figure 1).

Implementing SPREAD without knowing 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 require access to high performance computing resources. 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 assessed the effectiveness of different *h* values for approximating the true MMNND using the *N. crassa* dataset. We determined bootstrapped distributions of MMNND values for two different *h* values of 100 or 1000 by repeating SPREAD 1000 times. We varied the crossing-set size from $s_A = s_a = \{3, 4, 5, \ldots, 21\}$ To simplify the process, we used crossing-sets where $s_A = s_a$, but this is not a requirement of the SPREAD algorithm.

Plotting estimates of MMNND for both *h* values 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 decreases as the size of the crossing-sets increases. Furthermore, the max of the 1000 MMNND values for *h* = 100 is often less than for *h* = 1000; this difference becomes negligible as the number of crossing-sets increases. These 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, and using these *h*-values should be sufficient for most experiments. When using small crossing-set sizes, either *h* should be increased or the SPREAD algorithm repeated many times, and the crossing-set

with the max MMNND should be chosen from all iterations. With small or large crossing-set sizes the number of possible crossing sets is small so it may be possible to calculate an MMNND from all possible crossing-sets and determine the true MMNND.

Comparing SPREAD to SRS: 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 throughout cross characteristics (e.g. how reproductive success depends on the genetic or geographic distances between parents). To evaluate this hypothesis empirically, we used a simulated dataset of cross outcomes (reproductive output) and modeled relationships between these outcomes and the characteristics of crosses in crossing-sets generated from SPREAD or simple random sampling (SRS). The simulated cross outcomes were generated from "true" parameter values in a generalized linear model (GLM) relating cross characteristics to reproductive output. We repeated SPREAD or SRS 1000 times and refit the GLM to simulated data for each crossing-set for each method at each iteration. Figure 3 shows the bootstrapped distributions of parameter values generated from each sampling method and the "true" parameter values used to simulate the cross outcomes, along with the distributions of standard errors of the parameters. 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 are smaller and less variable compared to parameter values from crossing-sets generated through SRS (Figure 3b and Table 2). The SPREAD algorithm generates crossing-sets that provide greater extrapolative ability than crossing-sets generated by SRS, and will be very useful when trying to predict underlying

relationships between the characteristics of crosses and the reproductive outcomes of those crosses.

CONCLUSION

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 either the SPREAD algorithm or 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. R code for the implementation of the SPREAD algorithm is available at http://dx.doi.org/10.6084/m9.figshare.1180165 (ZIMMERMAN *et al.*). The SPREAD algorithm is not computationally intense; it can be used with populations of varying size across n-dimensional cross characteristic space. More broadly, a MMNND approach may be useful in any situation where over-dispersed sub-samples taken evenly across parameter space are needed.

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.

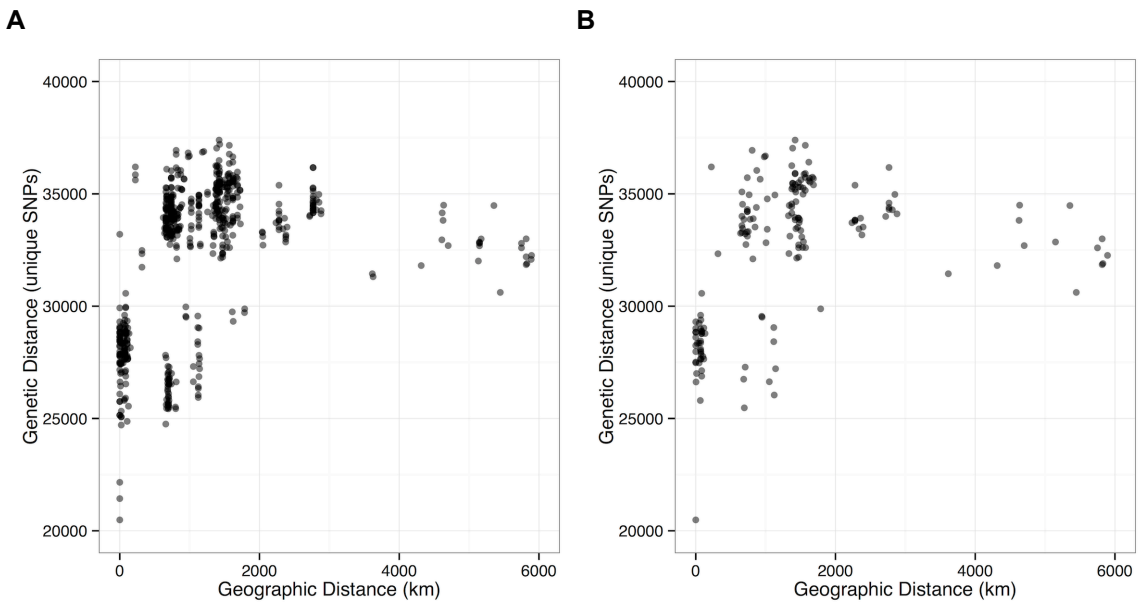

Figure 1 Implementation of SPREAD (code available for download at http://dx.doi.org/10.6084/m9.figshare.1180165 (ZIMMERMAN *et al.*)). Panel **a** shows all possible crosses between 24 mat-A and 24 mat-a strains; panel **b** shows the $s_A = s_a = 12$ crossing-set of *N. crassa* strains returned from 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. We used a previously published population genomics dataset - 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 (panel **a)**. We randomly sampled h = 1000 lists of s strains of each mating type from the set of all $\binom{\mathbb{A}}{s_A}$ and $\binom{\mathbf{a}}{s_a}$ strains without replacement, where $\mathbb{A}$ and $\mathbf{a}$ 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 12_A and 12_a 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 for use in an experiment (panel **b)**.

We implemented our algorithm and additional analyses in the R programming language (R Core Team, 2014) using the following packages: 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)

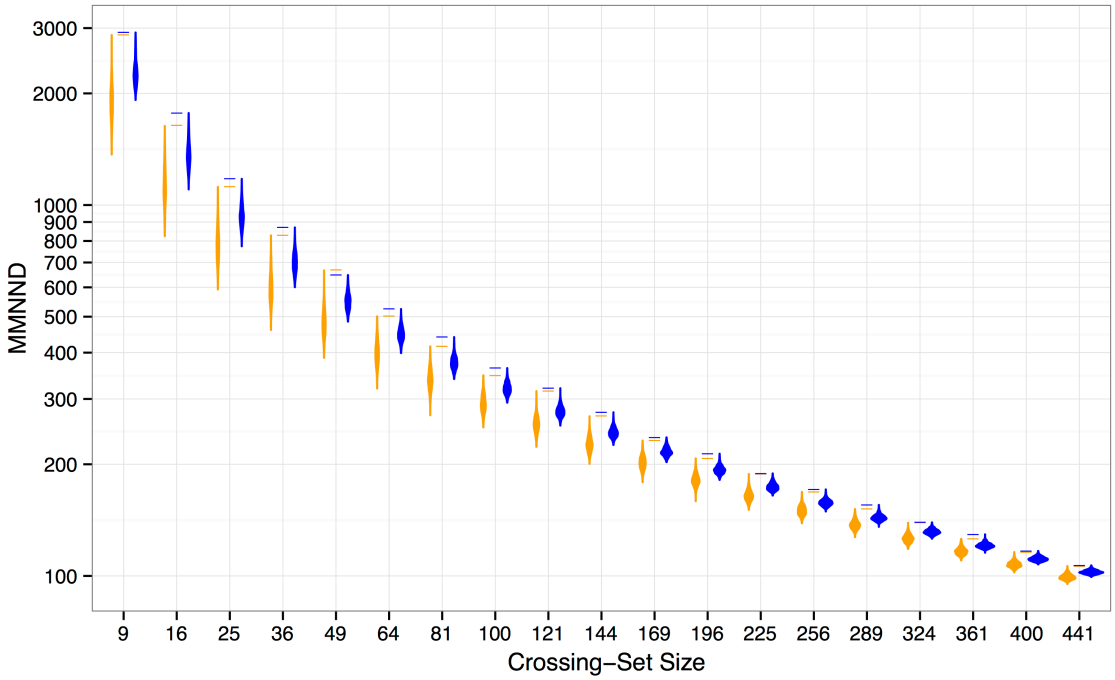

Figure 2 Effect of crossing-set size and *h* value on estimated MMNND values. With larger *h*-values, estimates of MMNND values increase. As crossing-set size increases MMNND values decrease because with more crosses the average distance between crosses decreases. Violin plots show the entire distributions of MMNND values plotted on a log scale. The two colors indicate the number of crossing-sets randomly sampled (orange: $h = 100$, blue: $h = 1000$). Dashes mark the maximum value of the distributions. The algorithm was implemented using the same method described in Figure 1 except we varied crossing-set size by implementing SPREAD for s_A and s_a, where s_A and s_a are integers from the set {3, 4, 5, …, 21} and $s_A = s_a$. 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.

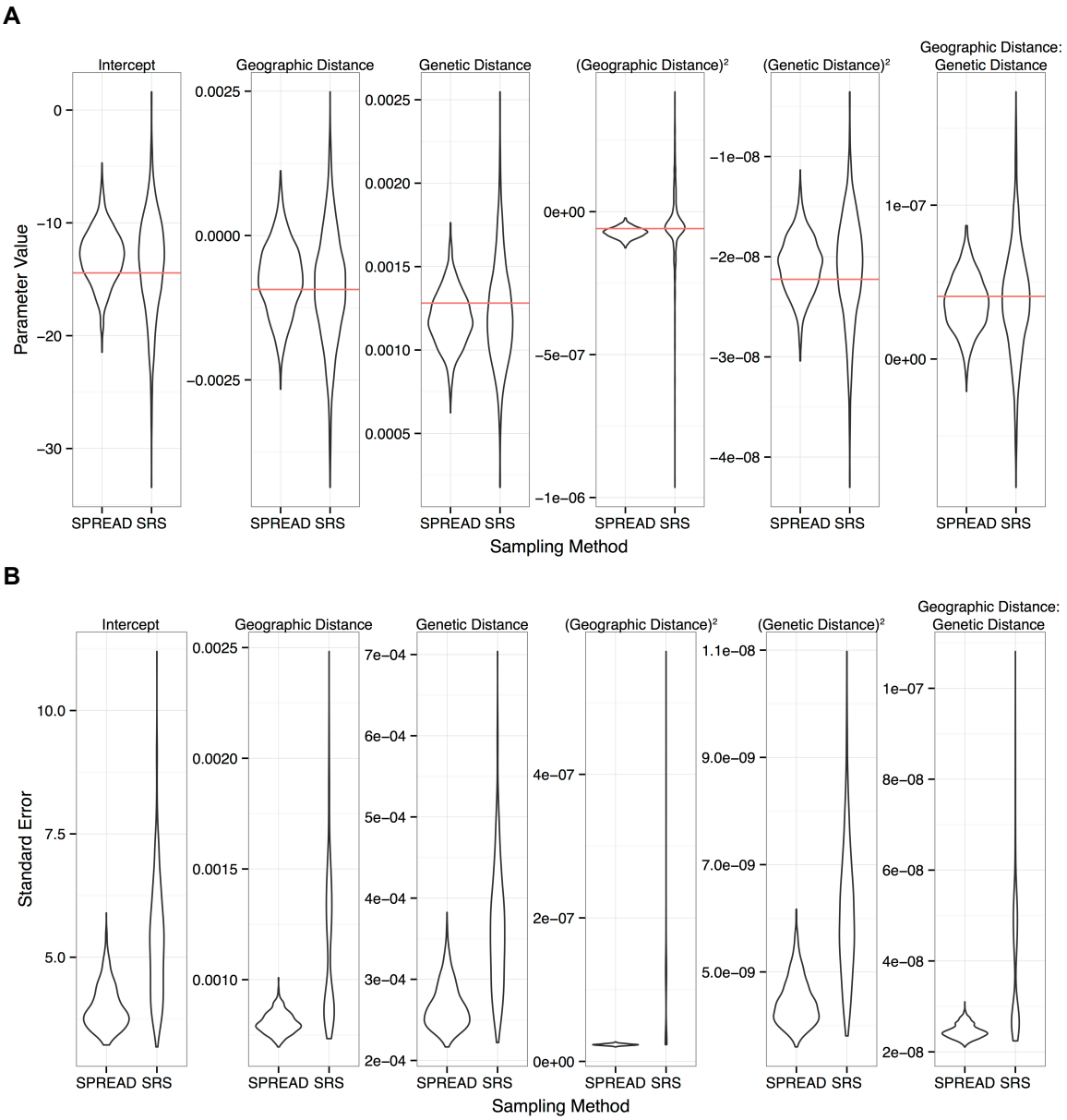

Figure 3 Comparisons of estimated parameter values and standard errors for 1000 crossing-sets generated either with SPREAD or SRS. Crossing-set size for both methods was with $s_A = s_a = 12$ and $h = 1000$ for SPREAD. Panel **a**: Violin plots showing the distribution of parameter values for all terms in the model. Red 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. To generate these

estimated parameter values and standard errors, and evaluate the effectiveness of SPREAD, we compared model fits to "true" values generated from simulated experimental data. Simulated experimental data, in the form of total ascospore counts, were generated for the entire $s_A = s_a = 24$ crossing-set.

First, we generated simulated data for all possible crosses using a generalized linear model based on unpublished empirical data. The model was *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 mean derived from the predicted experimental values and distribution parameters derived from the real-data model described above. "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. For SPREAD, the parameter values 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 using the same model described above.

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.

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

Table 2 Summary statistics of Standard Errors from 1000 model fits using SPREAD or SRS generated crossing-sets.

	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
$(\text{Genetic Distance})^2$	4.45E-09	6.09E-09	1.97E-19	1.55E-18
$(\text{Geographic Distance})^2$	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

LITERATURE CITED

BADDELEY A., TURNER R., 2005 Spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software **12**: 1–42.

CLARK P. J., EVANS F. C., 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 S. A., 2014 Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. Frontiers in Genetics **5**: 1–5.

ELLISON C. E., HALL C., KOWBEL D., WELCH J., BREM R. B., GLASS N. L., TAYLOR J. W., 2011 Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci USA **108**: 2831–2836.

FOURNIER D. A., SKAUG H. J., ANCHETA J., IANELLI J., MAGNUSSON A., MAUNDER M. N., NIELSEN A., SIBERT J., 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.

JUI P. Y., LEFKOVITCH L. P., 1992 Selecting the size of a diallel cross experiment. Theor. Appl. Genet. **85**: 21–25.

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 K. J. F., JANNINK J.-L., MCINTYRE L. M., 2005 Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity **96**: 139–149.

WANG X., CUMMING S. G., 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.

YU H., 2002 Rmpi: Parallel Statistical Computing in R. R News **2**: 10–14.

ZIMMERMAN K., ADDICOTT E., LEVITIS D., PRINGLE A., Mathematical formalization of the SPREAD algorithm for choosing fully factorial crossing sets. figshare. http://dx.doi.org/10.6084/m9.figshare.1180170

ZIMMERMAN K., ADDICOTT E., LEVITIS D., PRINGLE A., Functions for implementing the SPREAD algorithm for choosing fully factorial crossing sets. figshare. http://dx.doi.org/10.6084/m9.figshare.1180165