Aquaculture 510 (2019) 43–50

Contents lists available at ScienceDirect

Aquaculture

journal homepage: www.elsevier.com/locate/aquaculture

Bayesian hierarchical modeling of yield in incomplete diallel crosses of the Pacific oyster *Crassostrea gigas*

Xiaoshen Yin[*,1], Dennis Hedgecock

Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Keywords:
Diallel cross
Bayesian hierarchical model
Parent line selection
Pacific oyster
Crassostrea gigas

ABSTRACT

Identifying elite inbred parent lines that produce high-performing hybrid Pacific oyster seed requires diallel or factorial test crosses among lines, each acting as both a male and a female parent. Previously, we used the generalized linear model with fixed effects (i.e. GLM) to partition variance in yield, among hybrid families produced by a diallel cross, into causal genetic components—principally, general combining ability (*GCA*), specific combining ability (*SCA*), and reciprocal effect (*R*). However, GLM is extremely sensitive to missing information, which arises from loss of hybrid families for random environmental causes or from variation in the reproductive success of parent lines. To resolve this issue, we apply a Bayesian hierarchical model, which partitions yield variance into the familiar causal genetic components, while providing Bayesian shrinkage estimates incorporating the uncertainty of missing data. Our study suggests that correlation between observed yields and those predicted by the Bayesian model is high ($r^2 \geq 0.99$), for observed offspring, regardless of diallel completeness. Additionally, in analyses of complete diallel crosses, line-specific *GCA* rankings from GLM and Bayesian models are consistent for parent lines. Finally, comparing simulated complete and incomplete diallel datasets, we show the accuracy of predicted yield for families that are present and of parent-line ranking by *GCA* and the reliability of parent-line selection for double-cross hybrids, especially when non-parental lines (i.e. the four hybrid parents used to predict the yield of double-cross hybrids) are present. Our study demonstrates that the Bayesian hierarchical model performs as well as GLM in analyzing complete diallel crosses and can properly deal with incomplete diallel crosses for which GLM does not work. Therefore, the Bayesian hierarchical model is powerful in diallel analysis to select superior parent lines for producing high-yielding, hybrid, Pacific oyster seed.

1. Introduction

The Pacific oyster *Crassostrea gigas* shows remarkable heterosis (hybrid vigor) for yield and its underlying components, growth and survival (Hedgecock et al., 1995; Launey and Hedgecock, 2001; Pace et al., 2006; Hedgecock and Davis, 2007; Plough and Hedgecock, 2011). Yield of F_1 hybrids significantly exceed that of the better-yielding parent in 16 of 22 cases (Hedgecock and Davis, 2007), meeting Griffing's (1990) operational definition of heterosis, i.e. potence, $h_p = Q/L > 1.0$, where Q is twice the deviation of a hybrid from the mid-parent value and L is the absolute difference between the parent trait-values. The widespread phenomenon of "better-parent" heterosis in crosses of inbred lines of the Pacific oyster, when evaluated in aquaculture systems, suggests that production of hybrid cultivars by crossbreeding has great potential to improve oyster yield (Hedgecock and Davis, 2007).

Diallel crosses, factorial mating experiments using a set of inbred lines as both male and female parents, allow calculation of causal components of variance in yield—general combining ability (*GCA*), specific combining ability (*SCA*), and reciprocal effect (*R*). Estimates of these components of yield-variance, typically derived from the generalized linear model with fixed effects (GLM) for diallel designs (Griffing, 1956), allow selection of superior parent lines for commercial production of hybrid oysters. Hedgecock and Davis (2007) show that yields of the Pacific oyster increase with *GCA*, as expected (Langdon et al., 2003; de Melo et al., 2016), but that non-additive genetic components of yield variance, *SCA* and *R*, make substantial contributions to yield that are often larger than the contribution of *GCA*. However, Hedgecock and Davis (2007) also revealed limitations of the GLM for analyzing data from diallel crosses with missing data. In practice, diallel

[*] Corresponding author at: Department of Biological Sciences, University of Southern California, 3616 Trousdale Pkwy, Los Angeles, CA 90089, USA.
E-mail address: xyin.aqua@gmail.com (X. Yin).
[1] Current address: Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA.

https://doi.org/10.1016/j.aquaculture.2019.05.031
Received 2 April 2019; Received in revised form 12 May 2019; Accepted 13 May 2019
Available online 14 May 2019

crosses rarely conform to one of the classical designs, owing to reproductive failure or environmental factors (Hedgecock and Davis, 2007). Solutions to the problem of missing data include dropping parent lines, resulting in loss of power, collapsing incomplete diallel crosses to complete, pseudo-partial diallel crosses, diminishing the scope of inference about *SCA* and *R*, or imputation of missing data (Turner et al., 2018).

Here, we compare previous results obtained from GLM with results from an alternative, Bayesian hierarchical model for analyzing data from diallel crosses (Lenarcic et al., 2012). The Bayesian model partitions yield-variance into familiar parameters, i.e. additive, heterotic, epistatic, and parent-of-origin effects, while overcoming the difficulties posed by incomplete or imbalanced data and by data outliers. We find correspondences between the effects specified in the Bayesian hierarchical model and the classical components of yield-variance. We further explore the reliability of estimates of *GCA*, *SCA* and *R* provided by the Bayesian hierarchical model through simulations with two patterns of missing data, random loss of families and random loss of parent-line representation, mimicking the commonly observed reproductive failure of certain parent lines in diallel crosses of oysters (Lannan, 1980; Hedgecock and Davis, 2007). Finally, we focus on the reliability, in the face of missing data, of (1) ranking parent lines to identify those worthy of further consideration for production of single-cross hybrids and (2) identifying F_1 hybrids for production of double-cross hybrids. Jones (1918, 1922) promoted the use of double-cross hybrids in the early days of maize breeding, as a way of circumventing the handicap of producing F_1 hybrid seed on an inbred parent; double-cross hybrids likely present the same advantage for the Pacific oyster, since inbred oysters are small and produce inadequate numbers of eggs for commercial hatchery production.

As noted above, diallel crosses rarely produce the next generation of the parental lines, owing to inbreeding depression (Lannan, 1980; Launey and Hedgecock, 2001; Evans et al., 2004; Plough and Hedgecock, 2011). Since our main interest, here, is in the practical application of information from diallel crosses to the crossbreeding of superior hybrid cultivars, we focus on Griffing's (1956) Method 3 for analyzing a diallel cross that yields data on all $p(p - 1)$ F_1 hybrids from *p* parent lines. In this partial diallel, without inbred parents, we can estimate the extra-nuclear effects, *R*, causing differences between reciprocal hybrids.

2. Materials and methods

2.1. Data on yield from diallel crosses

We collected data on yield from six diallel crosses, which we name by birth year and experiment number (e.g. 01x1 stands for the first experimental cross set up in 2001; Table 1). Hedgecock and Davis (2007) previously reported analyses of data from four of these crosses; we have not previously reported analyses of data from the 12x1 and 12x2 crosses. We represent inbred parent lines by a number, which is an abbreviation of the full family name described by Hedgecock and Davis (2007), and first-generation hybrid offspring (F_1 families) by a sire × dam name (e.g. 2 × 10 stands for offspring with paternal parent from inbred line 2 and maternal parent from inbred line 10; Table 1).

We obtained yield (biomass) data at different phases in the production cycle for three diallel crosses, as described by Hedgecock and Davis (2007). Phase II is indoor replicated nursery culture in upwelling tubes; Phase III is outdoor replicated nursery culture in suspended, rotating seed cages; Phase IV is final grow out of adults in on-bottom cages. In simulations described below, we refer to Phase III oysters as juvenile and Phase IV oysters as adult.

With one exception (01x1-IV), the phenotype analyzed is mean live weight per individual oyster in a rearing unit at the end of a given phase of culture (i.e. total live weight of oysters per cage or tube divided by the count of live oysters). Since survival was generally high for all juveniles, mean live weight is a measure of biomass yield in this study. In 01x1-IV, we analyze total weight of oysters per cage. Altogether, we analyzed nine, partial or incomplete and four, complete diallel datasets (see Table 1 for diallel cross names, the number and names of parent lines, and culture phases for yield data). We identify complete diallel crosses that are embedded within larger, incomplete diallel crosses.

2.2. Statistical models

Traditionally, diallel crosses are analyzed by GLM (Griffing, 1956),

$$Y_{ijk} = g_i + g_j + s_{ij} + r_{ij} + e_{ijk} \tag{1}$$

where Y_{ijk} is mean live weight per individual, g_i is the general (additive) combining ability (*GCA*) of paternal parent *i*, g_j is the *GCA* of maternal parent *j*, s_{ij} is the specific (non-additive) combining ability (*SCA*) of hybrid $i \times j$, r_{ij} is the reciprocal effect (*R*), accounting for differences between reciprocal hybrids $i \times j$ and $j \times i$, when reciprocal hybrids are included, and e_{ijk} is experimental error.

In this study, we analyze diallel crosses, using a subset of the full Bayesian hierarchical model proposed by Lenarcic et al. (2012), who provide a detailed explanation of the BayesDiallel model and comparison with Griffing's method. One notable difference between BayesDiallel and conventional models of dominance is that BayesDiallel models heterosis as inbred-specific deviations from heterozygote-based predictions. In their terminology, we fit the “Babmvw” model, which incorporates interactions between pairs of parent lines. For hybrid families ($j \neq k$),

$$Y_{jki} = \mu + a_j + m_j + a_k - m_k + v_{jk} + w_{jk} + e_i \tag{2}$$

Table 1
Summary of diallel crosses and diallel datasets.

Diallel cross	Culture phase	Dataset code	No. parent lines	Parent lines
A) Incomplete diallel crosses				
01x1	III, IV	01x1-III, -IV	6	2, 10, 35, 38, 46, 51
01x4	III, IV	01x4-III, -IV	7	9, 28, 33, 35, 41, 46, 53
03x6	II, III	03x6-II, -III	9	20, 26, 35, 36, 45, 47, 51, 52, 92
03x8	IV	03x8-IV	7	3, 9, 19, 21, 40, 61, 94
12x1	III	12x1-III	10	8, 15, 19, 20, 21, 24, 32, 33, 34, 35
12x2	III	12x2-III	8	7–39, 16, 5, 26, 2–39, 30, 36, 45
B) Complete diallel crosses				
01x1	III, IV	01x1-IIIC, -IVC	5	2, 35, 38, 46, 51
03x6	II, III	03x6-IIC, -IIIC	5	26, 36, 45, 47, 92

Diallel cross is named by birth year and experiment number (e.g. 01x1 stands for the first experimental cross set up in 2001). Culture phase is defined as described by Hedgecock and Davis (2007): phase II is indoor replicated nursery culture in upwelling tubes; phase III is outdoor replicated nursery culture in suspended, rotating seed cages; phase IV is final grow out of adults in on-bottom cages. Dataset code is represented by diallel cross and culture phase (e.g. 01x1-III stands for data collected from diallel cross 01x1 at culture phase III). Cross codes followed by a “C” are complete diallel datasets; the rest are incomplete diallel datasets.

where Y_{jki} is mean live weight per individual, μ is the grand mean live weight per individual oyster in a diallel cross, a_j and a_k are additive (dosage) effects of parent lines, m_j and m_k are maternal effects of parent lines, v_{jk} is the family-specific symmetric effect for cross $j \times k$ (dam × sire), w_{jk} is the family-specific asymmetric effect for cross $j \times k$ (dam × sire), and e_i is error for an individual oyster i in the cross of maternal parent j × paternal parent k. For reciprocal hybrid families, $v_{jk} = v_{kj}$ and $w_{jk} = -w_{kj}$. For this BayesDiallel mixed model with six variance components, we estimate a set of priors on the variance components, which make the model Bayesian, and fit the model using a Markov Chain Monte Carlo (MCMC) Gibbs sampler with five chains, 3500 iterations, and a burn-in of 1000.

The genetic components, *GCA*, *SCA*, and *R*, in GLM correspond to analogous terms in the Bayesian hierarchical model, according to their biological interpretations. *GCA* in GLM is equal to the parental additive effects in the Bayesian hierarchical model. *SCA* corresponds to the family-specific symmetric effect. *R* corresponds to a combination of maternal and family-specific asymmetric effects for hybrid families. Posterior distributions and point estimates of these genetic effects and the average live weights of families were obtained from Bayesian analysis as implemented in BayesDiallel, a program running in R 3.0.2 (Lenarcic et al., 2012).

Finally, we compared observed and predicted yields by obtaining the regression equation and the coefficient of determination (r^2) between observed and predicted yields and between Bayesian- and GLM-predicted yields (Tables 2, 3).

2.3. Diallel cross simulation

To test the performance of BayesDiallel, we simulated diallel-cross datasets, using eq. (2) to obtain live weight per individual cage. Genetic effects (i.e. *a, m, v, w*) and μ in Eq. (2) are assumed to be normally distributed, so two parameters, mean and standard deviation, are needed to simulate μ and the genetic effects for each parent line and cross, the sum of which is simulated live weight per cage. Since yield should be positive, we took the absolute value of simulated yield for all cages. Two experimental factors modeled in these simulations are (i) the phase when data on live weight are collected (i.e. Phase) and (ii) the pattern of missing (or present) families (i.e. Pattern).

Table 2
Relationships among GLM-predicted, Bayesian-predicted, and observed yields in analyses of 5 × 5 complete diallel crosses.

Dataset	Regression equation	r^2
	GLM-predicted versus observed yield (*x*: observed yield; *y*: GLM-predicted yield)	
01x1-IIIC	$y = 1.0133x + 0.0561$	0.9974
01x1-IVC	$y = 0.9917x + 8.3825$	0.9975
03x6-IIC	$y = x - 0.061$	1.0
03x6-IIIC	$y = x - 0.0375$	1.0
	Bayesian-predicted versus observed yield (*x*: observed yield; *y*: Bayesian-predicted yield)	
01x1-IIIC	$y = 0.9833x + 0.0204$	0.9997
01x1-IVC	$y = 1.0292x - 86.783$	0.999
03x6-IIC	$y = 1.0066x - 0.0004$	0.9993
03x6-IIIC	$y = 0.7383x + 1.3485$	0.9904
	Bayesian-predicted versus GLM-predicted yield (*x*: Bayesian-predicted yield; *y*: GLM-predicted yield)	
01x1-IIIC	$y = 1.0304x + 0.0352$	0.9976
01x1-IVC	$y = 1.0206x - 77.394$	0.9963
03x6-IIC	$y = 0.9928x - 0.0161$	0.9993
03x6-IIIC	$y = 1.3416x - 1.7971$	0.9904

Dataset is coded by diallel cross and culture phase (e.g. 01x1-III stands for data collected from diallel cross 01x1 at culture phase III). Dataset codes followed by a "C" are complete diallel datasets. The data analyzed are average live weights per individual (g/individual) per cage, except for diallel dataset 01x1-IV, which are total biomass (g) per cage.

We used means and standard deviations for μ and genetic effects of seven inbred parent lines in Bayesian outputs for 12x1-III and 03x8-IV as means and standard deviations in simulations. We simulated μ and genetic effects (i.e. *a, m, v, w*) independently. We generated cage live weights in 7 × 7 diallel crosses at both juvenile (simulated based on 12x1-III Bayesian output) and adult (simulated based on 03x8-IV Bayesian output) life stages. Since we were not interested in inbred families, we removed all inbred families from simulated 7 × 7 diallel crosses. In total, we simulated 1000 complete diallel datasets (coded as CPLT) for each of the juvenile and adult life stages, each consisting of 42 hybrid families. We simulated the rearing of each family in 10 cages, yielding ten mean live weights per cage for each family.

Generally, two types of missing-family patterns are observed in practice: randomly missing families (symbolized as R) and randomly missing rows or columns, owing to loss of all hybrid families sharing a common parent line (symbolized as RRC; Lannan, 1980; Hedgecock and Davis, 2007; examples of these two patterns are in Fig. S1). For simulated R diallel crosses, 21 hybrid families (50% of hybrid families) were randomly removed from CPLT diallel crosses, generating incomplete diallel crosses consisting of 21 hybrid families. For simulated RRC diallel crosses, we randomly removed three rows or columns, producing incomplete diallel crosses consisting of 24 or 26 hybrid families.

In total, there are four Phase-by-Pattern combinations of simulated diallel datasets (i.e. Juvenile-R, Juvenile-RRC, Adult-R, and Adult-RRC). Since 1000 CPLT diallel datasets were simulated for each life stage, 1000 different 7 × 7 diallel datasets were generated under each Phase-by-Pattern combination by randomly removing different families (for R diallel crosses) and different rows or columns (for RRC diallel crosses) from CPLT diallel crosses. In total, 4000 incomplete diallel datasets were simulated, among which R and RRC diallel crosses consist of 210 and 240 (or 260) cages, respectively. All simulated diallel crosses and the R-code for simulating diallel-cross data are available upon request.

2.4. Parent-line ranking

Although Griffing (1956) takes the variance of *GCA* and *SCA* into consideration in selecting superior parent lines, we simply use *GCA* to explore the effect of missing data on the selection of top parent lines. For each complete diallel dataset (01x1-IIIC, 01x1-IVC, 03x6-IIC, 03x6-IIIC), we computed line-specific *GCA* estimates from GLM and Bayesian analyses. For each simulated incomplete diallel cross, we estimated line-specific *GCA*, using Bayesian analysis. We found the top three parent lines in each simulated complete diallel dataset and then counted the number of non-top-three parent lines identified in incomplete diallel datasets (mismatches). To test how the number of missing families influences the *GCA* ranking of a parent line, we recorded the difference in parent-line *GCA* ranks between each simulated complete diallel dataset and its corresponding incomplete diallel dataset (7 parent lines × 1000 simulations × 4 complete vs. incomplete pattern pairs). We then compared this difference in ranks against the number of missing families for the corresponding parent line for each Phase-by-Pattern combination, using one-way ANOVA, with the number of missing families as an independent variable and the rank difference as a dependent variable.

2.5. Double-cross hybrid parent selection

We simulated the generation of double-cross hybrids from crosses between two unrelated hybrid parents. The yield of double-cross hybrids can be predicted by the yield of hybrid families in a diallel cross according to Method B of Jenkins (1934), in which

$$Y_{(\mathrm{A}\times\mathrm{B})\times(\mathrm{C}\times\mathrm{D})} = (Y_{\mathrm{A}\times\mathrm{C}} + Y_{\mathrm{A}\times\mathrm{D}} + Y_{\mathrm{B}\times\mathrm{C}} + Y_{\mathrm{B}\times\mathrm{D}})/4 \tag{3}$$

where $Y_{(\mathrm{A}\times\mathrm{B})\times(\mathrm{C}\times\mathrm{D})}$ is the predicted yield of a double-cross hybrid AB × CD and $Y_{\mathrm{A}\times\mathrm{C}}$, $Y_{\mathrm{A}\times\mathrm{D}}$, $Y_{\mathrm{B}\times\mathrm{C}}$ and $Y_{\mathrm{B}\times\mathrm{D}}$ are the observed yields of

Table 3
Predictions of yield by Bayesian analysis for existing families in incomplete diallel crosses.

Diallel dataset	No. of families attempted	No. of families obtained	Observed yield (g/individual) (mean ± s.d.)	Predicted yield (g/individual) (mean ± s.d.)	Regression equation	r^2
01x1-III	30	25	1.17 ± 0.24	1.17 ± 0.22	$y = 0.98x + 0.02$	1.0
01x1-IV	30	25	3214 ± 475	3193 ± 424	$y = 1.04x\text{-}117.68$	0.998
01x4-III	42	21	0.95 ± 0.17	0.98 ± 0.14	$y = 0.97x + 0.03$	1.0
01x4-IV	42	21	25.55 ± 4.44	25.62 ± 3.25	$y = 0.81x + 4.89$	0.968
03x6-II	72	44	0.06 ± 0.02	0.06 ± 0.01	$y = x\text{-}0.0001$	0.999
03x6-III	72	44	5.18 ± 0.92	5.18 ± 0.57	$y = 0.77x + 1.19$	0.984
03x8-IV	42	26	60.04 ± 8.43	60.25 ± 6.74	$y = 0.94x + 3.77$	0.996
12x1-III	90	29	0.82 ± 0.27	0.84 ± 0.18	$y = 0.96x + 0.03$	1.0
12x2-III	56	25	1.75 ± 0.46	1.79 ± 0.34	$y = 0.96x + 0.07$	0.999

Observed and predicted yields are grand means across existing families. Linear regressions are Bayesian-predicted yield [dependent variable, y] on observed yield [independent variable, x] for existing families. Observed and predicted yields for diallel cross 01x1-IV are total biomass (g) per cage.

hybrid families A × C, A × D, B × C and B × D, respectively. Families A × B and C × D are called parental lines because they serve as parents for the double-cross hybrid; families A × C, A × D, B × C and B × D are called non-parental lines, because they are only used for predicting the yield of the double-cross hybrid and do not serve as parents.

In order to select the optimal parents for double-cross hybrids, we need accurate estimates of the yield of missing non-parental lines. We assesed the accuracy of predicted yields of missing non-parental lines, using the coefficient of correlation (r) between estimated yields in simulated complete diallel crosses and predicted yields in simulated incomplete (R or RRC) diallel crosses.

3. Results

3.1. Predictions of yield by GLM and Bayesian methods in analyses of data from diallel crosses

In analyses of data from complete diallel crosses, correlations between observed and predicted yields from GLM and Bayesian analyses are high for all hybrid families, with coefficients of determination (r^2) above 0.99 for both methods (Table 2). Predicted yields from both methods are also highly correlated. In analyses of data from incomplete diallel crosses, with as many as 67% of families missing, coefficients of determination (r^2) for linear regressions of Bayesian-predicted yield on observed yield for existing families range from 0.968 to 1 (Table 3).

3.2. Performance of BayesDiallel in simulated complete and incomplete diallel crosses

We use simulations based on real diallel datasets, representing juvenile and adult culture phases (12x1-III and 03x8-IV, respectively), to explore the performance of BayesDiallel under two patterns of missing data, randomly missing families (R) and randomly missing rows and columns (RRC). Each simulated, complete dataset generates R and RRC incomplete datasets for comparison. Across 1000 simulations for each of the four Phase-by-Pattern combinations, correlation coefficients (r) of predicted yields between simulated complete and incomplete diallel crosses are above 0.98 on average for existing families, but cluster from 0.4–1 for missing families (Fig. 1).

Predictions of line-specific *GCA* ranking by GLM and Bayesian analyses are consistent for all parent lines in complete diallel crosses (Fig. 2). Mean difference in rank of line-specific *GCA* between simulated complete and incomplete diallel crosses, for all parent lines, is affected only slightly but significantly by the number of missing families per parent line in the randomly missing, R-pattern ($F_{10,\ 6989} = 4.67$, $P < .0001$ for Adult-R; $F_{10,\ 6989} = 9.25$, $P < .0001$ for Juvenile-R). When entire rows or columns are missing (the RRC-pattern), then the number of missing families per parent line falls into four discrete classes (see Fig. S1) and makes a significant difference of about half a rank, when > 3 families are missing per parent line ($F_{3,\ 6996} = 214.26$, $P < .0001$ for Adult-RRC; $F_{3,\ 6996} = 116.9$, $P < .0001$ for Juvenile-RRC; Fig. 3).

3.3. Selection of superior parent lines for F_1 and double-cross hybrids

We also use simulations to explore the accuracy and precision of identifying superior parent lines in analyses of incomplete diallel crosses. We, first, rank parent lines in simulations of complete diallel sets by line-specific *GCA*. Next, we find the top three parent lines in each simulated, incomplete diallel dataset and count how many of these match the top-three parent lines identified in the paired, complete diallel dataset. Across 1000 simulations of each of four types of incomplete diallel datasets (Juvenile-R, Juvenile-RRC, Adult-R and Adult-RRC), top-three parent lines are matched to those in the paired, simulated, complete diallel dataset, an average of 2.24 ± 0.02, 2.20 ± 0.02, 2.76 ± 0.01, and 2.55 ± 0.02 times, respectively (Table 4).

4. Discussion

4.1. Comparison of GLM and Bayesian analyses

Diallel analysis allows partitioning of variance in yield into the genetic components of general combining ability (*GCA*), specific combining ability (*SCA*) and reciprocal effect (*R*). In our study of complete diallel-cross datasets, GLM and Bayesian hierarchical models explain > 99% of the variance in observed yields of young seed, juveniles, and adults (culture Phases II to IV; Table 2). This indicates that both GLM and Bayesian analysis are powerful tools for complete diallel analysis, regardless of which life stage furnishes the yield data. However, in practice, unanticipated missing information makes it hard to extract information from a diallel cross, using traditional GLM methods. Recent development of imputation methods allows for analyses of crosses with missing data (Turner et al., 2018), but these methods are still quite cumbersome, requiring extra computational steps. BayesDiallel, on the other hand, provides a straightforward means for conducting Bayesian analyses of incomplete diallel-cross data. Our simulations show that, even when up to half of the hybrid families are lost from a full diallel cross, Bayesian analysis can still effectively estimate the genetic components *GCA*, *SCA* and *R*, and accurately predict yield

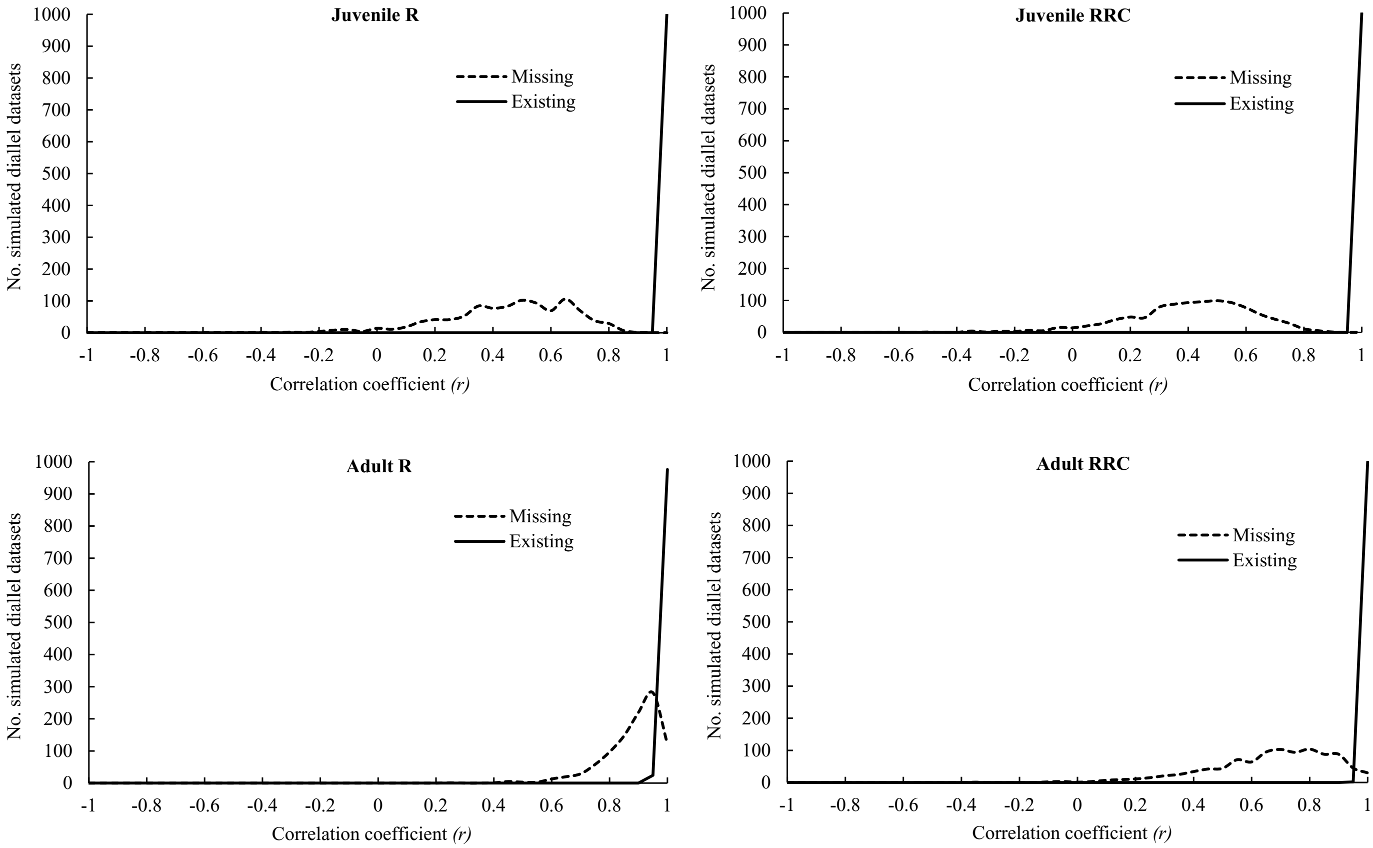

Fig. 1. Density plots of correlation coefficients (*r*) of predicted yields between simulated complete and incomplete (R and RRC) diallel crosses for missing and existing F_1 hybrid families. The "No. simulated diallel datasets" is the number of diallel datasets with a certain *r*. Juvenile and Adult represent diallel crosses simulated based on 12x1-III and 03x8-IV Bayesian outputs, respectively; R and RRC stand for two missing patterns.

for existing families (Fig. 1; Fig. S2). At the same time, these simulations suggest caution in making inferences about the yield of missing families, since the accuracy and precision of these estimates are rather poor (Fig. 1); however, this limitation may not be of practical concern for crossbreeding.

4.2. Line-specific GCA ranking and parent line selection

We selected top parent lines according to their line-specific *GCA* (Griffing, 1956). Comparing ranks of line-specific *GCA* estimates from GLM and Bayesian analysis, we find that the two methods yield consistent rankings for all parent lines in four, 5 × 5 complete diallel datasets. Therefore, Bayesian analysis provides reliable information for selecting superior parent lines in complete diallel crosses.

In our simulations, an average of 2.2 to 2.8 of three top parent lines are correctly selected according to line-specific *GCA* rank, regardless of missing pattern or life stage (Table 4). This result suggests that the Bayesian hierarchical model is reliable for estimating *GCA* and useful for selecting superior parent lines for further development and crossbreeding. Parent-line selection is less accurate for RRC than for R diallel crosses, because less information is available for a specific parent line, when, owing to reproductive failure of some parent lines, entire rows or columns are missing (Table 4). In contrast, when families are missing randomly, owing to random environmental factors, the loss of information does not appear to have a severe impact on the analysis of the diallel cross and on picking superior parent lines.

For each missing pattern, parent-line selection appears to be more accurate for the Adult than for the Juvenile stage (Table 4), but this is likely attributable to different signal-to-noise ratios in the underlying datasets (i.e. variance among families divided by averge variance within family). Mean signal-to-noise ratios for diallel crosses Juvenile-R, Juvenile-RRC, Adult-R and Adult-RRC are 1.66, 1.61, 6.49 and 6.15, respectively. This suggests that decreased signal-to-noise ratios in yield data collected from younger life stages may decrease the accuracy of top parent-line selection, but signal-to-noise ratio can be increased by setting up more replicate cages for each family. Meanwhile, within a certain range, the number of missing families of a parent line also affects the accuracy of line-specific *GCA* ranking for that parent line. For diallel crosses with random losses of families, the line-specific *GCA* ranking tends to be only slightly less accurate as the number of missing families per parent line increases from one to ten (rank differences, 0.377 vs. 0.613, respectively). For diallel crosses with randomly missing parent-line representation, the *GCA* rank difference is much smaller when only three families are missing for a parent line (Fig. 3). Regardless of patterns and stages of diallel crosses, the predicted (CPLT-R) or median (CPLT-RRC) absolute line-specific *GCA* rank difference of parent lines is around or smaller than 1 (Fig. 3), demonstrating accurate line-specific *GCA* rankings based on the Bayesian hierarchical model in the face of missing data.

4.3. Double-cross hybrids

Since inbred oysters are small and produce inadequate numbers of eggs, commercial production of hybrid oysters can borrow a method

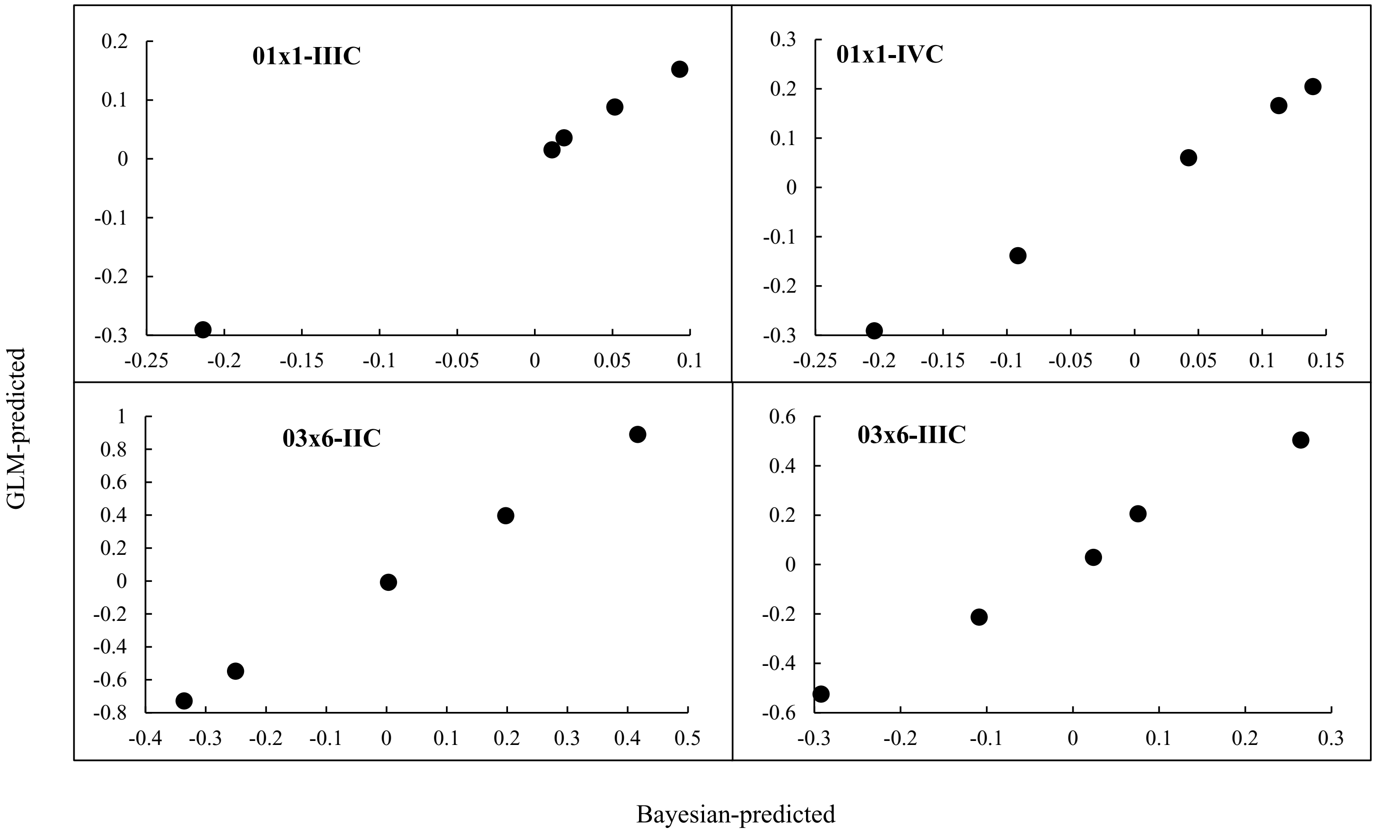

Fig. 2. GLM-predicted vs. Bayesian-predicted line-specific *GCA*. GLM-predicted and Bayesian-predicted line-specific *GCA* refer to predications based on the generalized linear model with fixed effects and the Bayesian hierarchical model, respectively.

from the early days of hybrid corn breeding, creating double-cross hybrid families by crossing two, unrelated F_1 hybrid parents. This strategy permits use of robust, highly fecund, F_1 hybrid females for commercial hatchery crosses. We may select F_1 parents for double-cross hybrids, by estimating double-cross hybrid performance from the yields of F_1 non-parental lines (Jenkins, 1934). Analyses of both real and simulated incomplete diallel crosses demonstrate that the Bayesian hierarchical model can accurately predict the yield of existing F_1 families (Table 3; Fig. 1). Therefore, we may be confident in parents selected for double-cross hybrids, when all non-parental lines are present in a diallel cross, because the estimate of double-cross hybrid yield is accurate.

However, some non-parental lines are lost in diallel crosses in practice. In this case, we need to be cautious in selecting superior parents for double-cross hybrids. Predictions of yields for the Adult stage appear to be better than those for the Juvenile stage, and at the Adult stage, yield is more accurately estimated for R than for RRC diallel cross (Fig. 1). Again, the better predictions for the Adult stage may be driven by a larger signal-to-noise ratio of yield data collected at the later life stage. Thus, as in parent-line selection for F_1 hybrids, the design of diallel crosses should aim to increase the signal-to-noise ratio of yield data. Moreover, predicted yields of F_1 parents for making double-cross hybrids and estimating their yields appear to be more accurate in incomplete diallel crosses with randomly missing families driven by environmental factors than in those with missing parent-line representation driven by reproductive failures.

Based on our analyses of simulated diallel crosses, the overall prediction of yields for missing F_1 families does not seem to be very reliable for selecting parents for double-cross hybrids (Fig. 1). The poor prediction of yield for missing F_1 families is largely attributable to inaccurate estimation on *SCA* and *R* (Fig. S2), suggesting that, in addition to *GCA, SCA* and *R* are also important factors to consider in parent line selection for both F_1 and double-cross hybrids. Improving the accuracy of predicting *SCA* and *R* could make parent-line selection more reliable.

In practice, parent-line selection should be more promising for double-cross hybrids. In our study, we assume that ~40% of families are missing in a simulated diallel cross, but a higher percentage of families often survive in reality. Availability of more unrelated hybrid families in a diallel cross can generate a sufficient number of double-cross hybrid families without any non-parental line missing. In addition, a larger number of surviving families in a diallel cross can decrease the number of missing non-parental lines for potential double-cross hybrids, thus improving the accuracy of estimates on the double-cross hybrid yield. Even if such a high proportion (i.e. ~40%) of families were missing in practice, one could reduce the dimension of the diallel crosses for analysis, in which case the accuracy of predicted yields would improve. Therefore, compared to our simulated study, selecting parents for setting up double-cross hybrids in practice should be more reliable.

5. Conclusions

By partitioning variance in yield among families in a diallel cross, regardless of its completeness, into *GCA*, *SCA* and *R* components, the Bayesian hierarchical model provides a powerful analytical means for selecting superior parent lines to improve the yield of hybrid Pacific

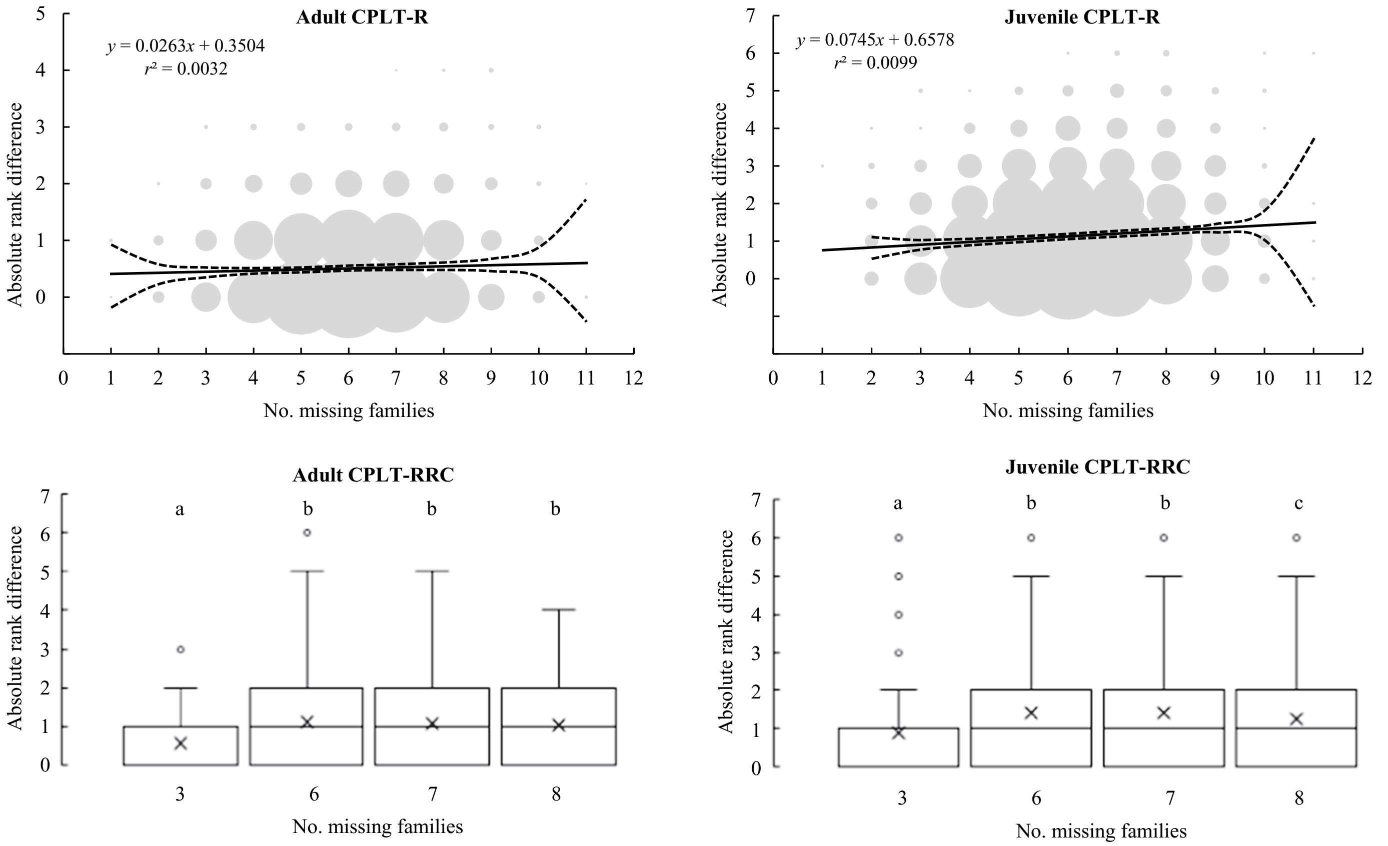

Fig. 3. Absolute difference in rank of line-specific *GCA* between simulated complete (CPLT) and incomplete (R and RRC) diallel crosses versus the number of missing families per parent line. Size of bubbles represents the number of observations with specific no. missing families and absolute rank difference. For each Phase-by-Pattern combination, 7000 parent lines are simulated. CPLT-R: the rank of line-specific *GCA* for a parent line in diallel cross CPLT minus the rank of line-specific *GCA* for the same parent line in diallel cross R. CPLT-RRC: the rank of line-specific *GCA* for a parent line in diallel cross CPLT minus the rank of line-specific *GCA* for the same parent line in diallel cross RRC. Juvenile and Adult represent diallel crosses simulated based on 12x1-III and 03x8-IV Bayesian outputs, respectively. In CPLT-R plots, solid and dashed lines indicate the line of best fit and 95% confidence interval, respectively. In CPLT-RRC plots, letters above boxplots indicate which mean rank differences (×), among the four categories of missing families, are significantly different.

Table 4

The number of matches and mismatches of the top three parent lines (by *GCA* ranking) between simulated complete (CPLT) and incomplete (R and RRC) diallel crosses.

No. mismatches per simulation/ summary statistics	CPLT vs Juvenile-R	CPLT vs Juvenile-RRC	CPLT vs Adult-R	CPLT vs Adult-RRC
0	339	313	763	570
1	560	580	237	409
2	100	102	0	21
3	1	5	0	0
Total no. mismatches	763	799	237	451
Total no. parent lines to keep	3000	3000	3000	3000
Average no. mismatches	0.76	0.8	0.24	0.45
Average no. matches	2.24 ± 0.02	2.2 ± 0.02	2.76 ± 0.01	2.55 ± 0.02
Proportion of matches	0.75	0.73	0.92	0.85

Total no. mismatches: the total number of mismatches in top three parent lines between simulated complete and incomplete diallel crosses across 1000 simulated diallel datasets of the same Phase-by-Pattern combination. Total no. parent lines to keep: the total number of top three parent lines that need to be kept across 1000 simulated diallel datasets of the same Phase-by-Pattern combination. Average no. mismatches per simulation (i.e. per top three parent lines) = Total no. mismatches / 1000 simulations. Average no. matches per simulation (i.e. per top three parent lines) = Total no. matches / 1000 simulations. Average no. matches per top three parent lines ± s.e. (of no. of matches per top three parent lines across 1000 simulations).

oysters. Line-specific *GCA*, predicted by the Bayesian hierarchical model, is a good parameter for correctly identifying top parent lines for further development and crossbreeding. Increasing replication and signal-to-noise ratios, while reducing the number of missing families, especially by properly conditioning parent lines to prevent loss of rows and columns in a diallel, are promising approaches to increasing the accuracy of predicting inbred and hybrid parent-line performance.

Acknowledgements

We gratefully acknowledge support from the NOAA Sea Grant Aquaculture Research Program 2010 [Project Number R/AQ-132NSI] administered by the California Sea Grant Program, under NOAA omnibus award NA10OAR4170060. We would also greatly thank Taylor Shellfish Farms for assistance with field maintenance of oysters.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aquaculture.2019.05.031.

References

de Melo, C.M.R., Durland, E., Langdon, C., 2016. Improvements in desirable traits of the Pacific oyster, *Crassostrea gigas*, as a result of five generations of selection on the West Coast, USA. Aquaculture 460, 105–115. https://doi.org/10.1016/j.aquaculture.2016.04.017.

Evans, F., Matson, S., Brake, J., Langdon, C., 2004. The effects of inbreeding on performance traits of adult Pacific oysters (*Crassostrea gigas*). Aquaculture 230, 89–98. https://doi.org/10.1016/j.aquaculture.2003.09.023.

Griffing, B., 1956. Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9, 463–493.

Griffing, B., 1990. Use of a controlled-nutrient experiment to test heterosis hypotheses. Genetics 126, 753–767.

Hedgecock, D., Davis, J.P., 2007. Heterosis for yield and crossbreeding of the Pacific oyster *Crassostrea gigas*. In: Aquaculture, Supplement: Genetics in Aquaculture IX 272, pp. S17–S29. https://doi.org/10.1016/j.aquaculture.2007.07.226. Supplement 1.

Hedgecock, D., McGoldrick, D.J., Bayne, B.L., 1995. Hybrid vigor in Pacific oysters: An experimental approach using crosses among inbred lines. In: Aquaculture, Genetics in Aquaculture V Proceedings of the Fifth International Symposium on Genetics in Aquaculture. vol. 137. pp. 285–298. https://doi.org/10.1016/0044-8486(95)01105-6.

Jenkins, M.T., 1934. Methods of estimating the performance of double crosses in corn 1. Agron. J. 26, 199–204. https://doi.org/10.2134/agronj1934.00021962002600030004x.

Jones, D.F., 1918. The effect of inbreeding and crossbreeding upon development. Proc. Natl. Acad. Sci. U. S. A. 4, 246–250.

Jones, D.F., 1922. The productiveness of single and double first generation corn hybrids. J. Am. Soc. Agron. 14, 242–252.

Langdon, C., Evans, F., Jacobson, D., Blouin, M., 2003. Yields of cultured Pacific oysters *Crassostrea gigas* Thunberg improved after one generation of selection. Aquaculture 220, 227–244. https://doi.org/10.1016/S0044-8486(02)00621-X.

Lannan, J.E., 1980. Broodstock management of *Crassostrea gigas*: I. genetic and environmental variation in survival in the larval rearing system. Aquaculture 21, 323–336. https://doi.org/10.1016/0044-8486(80)90067-8.

Launey, S., Hedgecock, D., 2001. High genetic load in the Pacific oyster *Crassostrea gigas*. Genetics 159, 255–265.

Lenarcic, A.B., Svenson, K.L., Churchill, G.A., Valdar, W., 2012. A general Bayesian approach to analyzing diallel crosses of inbred strains. Genetics 190, 413–435. https://doi.org/10.1534/genetics.111.132563.

Pace, D.A., Marsh, A.G., Leong, P.K., Green, A.J., Hedgecock, D., Manahan, D.T., 2006. Physiological bases of genetically determined variation in growth of marine invertebrate larvae: a study of growth heterosis in the bivalve *Crassostrea gigas*. J. Exp. Mar. Biol. Ecol. 335, 188–209. https://doi.org/10.1016/j.jembe.2006.03.005.

Plough, L.V., Hedgecock, D., 2011. Quantitative trait locus analysis of stage-specific inbreeding depression in the Pacific oyster *Crassostrea gigas*. Genetics 189, 1473–1486. https://doi.org/10.1534/genetics.111.131854.

Turner, S.D., Maurizio, P.L., Valdar, W., Yandell, B.S., Simon, P.W., 2018. Dissecting the genetic architecture of shoot growth in carrot (*Daucus carota* L.) using a diallel mating design. G3 Genes Genomes Genet. 8, 411–426. https://doi.org/10.1534/g3.117.300235.