QInfoMating: Sexual Selection and Assortative Mating Estimation Software

A. Carvajal-Rodríguez

Centro de Investigación Mariña (CIM), Departamento de Bioquímica, Genética e Inmunología. Universidade de Vigo, 36310 Vigo, Spain.

email: acraaj@uvigo.es

Running headline: QInfomating

Acknowledgments: This work is dedicated to my sister Paci. I thank E. Rolán-Alvarez for comments on the manuscript.

Data availability: The datasets generated and/or analysed during the current study are available in the DRYAD repository, http://doi.org/10.5061/dryad.h214h8t

Conflict of interest and author contribution statements: The authors declare that they have no competing interests. ACR is the sole author of the work.

Abstract

Background: Sexual selection theory is a multifaceted area of evolutionary research that has profound implications across various disciplines, including population genetics, evolutionary ecology, animal behavior, sociology, and psychology. It explores the mechanisms by which certain traits and behaviors evolve due to mate choice and competition within a species. In the context of this theory, the Jeffreys divergence measure, also known as population stability index, plays a key role in quantifying the information obtained when a deviation from random mating occurs for both discrete and continuous data.

Despite the critical importance of understanding mating patterns in the context of sexual selection, there is currently no software available that can perform model selection and multimodel inference with quantitative mating data. Recognizing this gap, I have developed QInfoMating which provides a comprehensive solution for analyzing and interpreting mating data within the framework of sexual selection theory.

Results: The program QInfoMating incorporates a user-friendly interface for performing statistical tests, best-fit model selection, and parameter estimation using multimodel inference for both discrete and continuous data. A use case is presented with real data of the species Echinolittorina malaccana.

Conclusions:

The application of information theory, model selection, and parameter estimation using multimodel inference are presented as powerful tools for the analysis of mating data, whether quantitative or categorical. The QInfoMating program is one of the few (if not the only) tools designed to perform this type of analysis.

Contact: acraaj@uvigo.es

Keywords: assortative mating, sexual selection, Jeffreys divergence, population stability index, Echinolittorina malaccana.

Introduction

By this point, there is no doubt about the importance of sexual selection theory and its impact in areas such as population genetics, evolutionary ecology, animal behavior, and even sociology and psychology. There are several strategies for measuring the impact of sexual selection, depending on the field of study and whether the available data pertain to discrete or continuous traits [1–3]. In previous works [2–4], it has been demonstrated

that the symmetric Kullback-Leibler divergence (also Jeffreys divergence or population stability index), noted as J_{PTI} in [3] corresponds to the increase in information when mating is not random. This divergence can be broken down additively into one component that measures sexual selection and another that measures assortative mating, and these statistics can be applied both to discrete and continuous data. Therefore, J_{PTI} unifies the measure of sexual selection for both discrete and continuous traits [2].

The program QInfoMating is a major update (see below) to the Infomating program [3, 4] and incorporates the new quantitative methods developed in [2]. The previous version (Infomating version ≤ 0.4) have already been utilized in empirical studies [5, 6]. While there are other software options available for computing sexual selection and assortative mating from mating tables, such as JMating [7], to the best of my knowledge, QInfoMating is the only software for applying model selection theory to mating data, enabling the estimation of sexual selection and assortative mating for both discrete and quantitative data.

Implementation

The backend of the program is implemented in C++11, while the interface is Python 3. Therefore, QInfomating can be run via the user interface (Figure 1-A) or directly from the console under Windows, Linux and macOS. Detailed instructions for use are provided in the program manual (https://acraaj.webs.uvigo.es/InfoMating/QInfoMating_v1.0_Manual.pdf), specifying input formats, the meaning of parameters, models, and result files. In summary, QInfoMating offers the capability to analyze both continuous and discrete data for sexual selection and assortative mating. It begins by performing relevant statistical tests to detect sexual selection and assortative mating and then identifies the best-fitting

model. The program employs multi-model inference techniques to estimate parameter values.

Difference of QInfoMating with the previous version (Infomating)

Unlike the previous version (Infomating), which lacked a user-friendly interface and only analyzed discrete data, QInfomating can be run via the user interface (Figure 1-A) and now accepts continuous data as input (see the format specification in the program manual). When continuous data is provided, the new statistical tests for sexual selection and assortative mating detection, as developed in [2], are performed.

Sexual selection tests

Briefly (see ref. [2] for a more detailed explanation), we assume that traits *X* and *Y* are normally distributed and the marginal density function for the trait in females within matings is $f_1(x) \sim N(\mu_1, \sigma_1^2)$, and in the population is $f(x) \sim N(\mu_x, \sigma_x^2)$ then, the test for sexual selection in females is

$$J_{S1} = \frac{1}{2}\left(\frac{\Phi_1^2+1}{\Phi_1} + \frac{\Phi_1+1}{\Phi_1}\frac{(\mu_1-\mu_x)^2}{\sigma_x^2} - 2\right)$$

where $\Phi_1 = \sigma_1^2/\sigma_x^2$.

and similarly, the marginal density function for the trait in males within matings is $f_2(y) \sim N(\mu_2, \sigma_2^2)$ and in the population is $g(y) \sim N(\mu_y, \sigma_y^2)$ then, the test for sexual selection in males is

$$J_{S2} = \frac{1}{2}\left(\frac{\Phi_2^2+1}{\Phi_2} + \frac{\Phi_2+1}{\Phi_2}\frac{(\mu_2-\mu_y)^2}{\sigma_y^2} - 2\right)$$

where $\Phi_2 = \sigma_2^2/\sigma_y^2$

and $J_{PSS} = J_{S1} + J_{S2}$.

For a random sample of n matings, nJ_{S1} and nJ_{S2} are asymptotically χ^2 distributed with 2 degrees of freedom [8, 9] under the null hypothesis that the females (males) within matings are distributed with mean μ_x (μ_y) and variance σ_x^2 (σ_y^2) i.e. absence of sexual selection.

If we want to test only the mean value of the female trait X in the mating sample, we set $\Phi_1=1$ so that

$$J_{S1\Phi_1=1}=\frac{(\mu_1-\mu_x)^2}{\sigma_1{}^2}=SSI_1{}^2$$

where SSI_1 is the standardized sexual selection intensity index [10–12].

And similarly for males

$$J_{S2\Phi_2=1}=\frac{(\mu_2-\mu_y)^2}{\sigma_2{}^2}=SSI_2{}^2.$$

Then with a sample size of n matings we have $(nJ_{S1\Phi=1})^{0.5} \sim t_{n-1}$ and $(nJ_{S2\Phi=1})^{0.5} \sim t_{n-1}$.

For large sample sizes both types of tests (χ^2 or t) work similarly, for a smaller sample size the χ^2 test might be slightly liberal. More generally, the χ^2 test should also detect differences in the variances, i.e., stabilizing (variance reduction) or diversifying (variance increasing) selection, or combinations of these forms of selection with directional selection (differences in means).

Assortative mating test

For assortative mating, assuming the quantitative traits we are measuring are normally distributed and jointly normally distributed within mating pairs, the information measure of assortative mating is a function of the square of correlation coefficient that takes values within the set of non-negative real numbers [0, +∞), specifically (see ref [2] for a more detailed explanation):

$$J_{PSI}=\int_{-a}^{+a}\int_{-a}^{+a}(q'(x,y)-f_1(x)f_2(y))\log\frac{q'(x,y)}{f_1(x)f_2(y)}dxdy=\frac{\rho^2}{1-\rho^2}$$

where $q'(x, y)$ is the mating probability density for pairs with values in the infinitesimal interval ([x, $x + dx$], [y, $y + dy$]) and $f_1(x)$ and $f_2(y)$ are the marginal density functions for the trait within matings in females and males, respectively [2].

For a random sample of n matings, nJ_{PSI} is asymptotically χ^2 distributed with 1 degree of freedom under the null hypothesis that the mating probability distribution is equal to the product of the female and male marginal distributions within matings [9].

Class frequency distribution

Once the statistical tests have been carried out, model selection is only executed if any of the statistical tests yield a significant result. Following this, the program automatically calculates the frequency distribution (classes) for the continuous data and proceeds with multi-model estimation. The algorithm to obtain the frequency distribution is detailed in the program manual. But it consists essentially of:

1- The range of the data set is calculated (maximum value *Max* - minimum value *Min*).

2- The formula that relates number of classes K and sample size n is [13]

$$K[\ln(K)+0.5777]+\frac{1}{2}\leq\frac{n}{20\ln(K)}$$

where we conservatively assume a sample size more than one order of magnitude larger than that required for the uniform frequency case (details in the program manual).

3- The width w of each class is computed as w = (range / K).

4- Make the classes. The first class is always [*Min*, *Min* + w) and the last class is [x, *Max* + w) where x is the maximum of the range of the previous class.

5- Compute the frequency for each class. Check that it is represented in each sex, if not, decrease *K* and repeat from 3.

Set of models

There are also differences in the set of models considered for discrete and continuous cases. In the discrete case, there is a default set of models that can be modified by the user. Conversely, for the continuous case, the program, rather than the user, selects models based on the statistical significance of various tests. Specifically, if there is a statistically significant pattern of assortative mating but no sexual selection, the program uses a set of models that incorporate mate choice and bias in conjunction with random mating and the saturated model, but it does not include competition models. On the other hand, if there is a statistically significant pattern of both assortative mating and sexual selection in both sexes, QInfoMating includes double-effect models among the models to be tested. These models are not included by default when the data is discrete.

In any case, QInfoMating still allows us to generate a mating table from the quantitative data, enabling us to create an input file for analysis in discrete mode, employing the default set of models or those selected by the user.

Results

Real data use case

Echinolittorina malaccana, a marine intertidal snail commonly found on rocky shores in the Indo-West Pacific region [14], exhibits mate selection behaviors. Males have shown a clear preference for females slightly larger than themselves [11]. In a subsequent study conducted by [6], the continuous trait of size (specifically, shell length) was transformed into categorical variables by creating four distinct size classes. This transformation

allowed for the creation of a mating table (see Table 1), which enabled the utilization of the multimodel inference techniques available in the previous version of the program (Infomating). This approach provided a more profound insight into the intricate dynamics of mate choice. The analysis ultimately confirmed that the most suitable model for mate choice in Echinolittorina malaccana involves a size bias, resulting in both assortative mating and the presence of sexual selection favoring larger individuals.

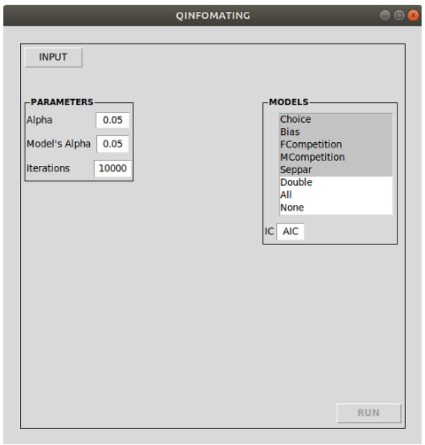
In a recent study [2], the same onshore data, originally presented in [6], and accessible via [15], was directly used without converting the continuous shell length trait into discrete classes. This approach aimed to demonstrate the computation of Jeffreys divergence indices [3] for quantitative traits. The research revealed that the statistical tests for the quantitative data yielded similar conclusions as the previous analysis using the statistical tests for discrete data. However, the quantitative data lacked the necessary structure for applying the model selection methodology.

In this current study, we will leverage QInfoMating to conduct a quantitative analysis of the same data set. The program will apply the quantitative statistical tests and automatically transform data into classes to facilitate model selection and multi-model inference. This analysis aims to compare the results with those obtained in [6]. For consistency, we will employ the Akaike Information Criterion (AIC), the same criterion used in [6], and additionally, the more conservative Bayesian Information Criterion (BIC) to showcase the software's versatility and utility.

A CSV file containing shell length data for individual snails and their mate pairs served as the input for QInfoMating. Although the original study by Lau et al. in [6], used four classes, QInfoMating segmented the data into three classes (Table 2), before proceeding with the model selection process.

If Figure 1-B the best-fit model from Lau et al. (B-7P, bias with 7 parameters model) and the two best obtained by QInfoMating (D-3P, double 3 parameters, and B-5P, bias 5 parameters models) are presented. The D-3P model is a model with double effect i.e., generates sexual selection and assortative mating patterns. From Figure 1-B in the D-3P model we appreciate that the choice parameters (main diagonal) are c_1 and c_3. The parameter c_1 can also be considered a male sexual selection parameter jointly with c_2 since they change the marginal propensity for the second class males (second column). However, from the point of view of biased choice, the parameter c_2 can be seen as a bias (subdiagonal) in a model in which individuals in the shorter class have more difficulty mating. In fact, the second best model was B-5P, which is the equivalent model to B-7P when there are three classes instead of four.

A



B

B-7P model

C_1	1	1	1
B_1	C_2	1	1
1	B_2	C_3	1
1	1	B_3	C_4

Best model in Lau et al.

0 ± 0	1 ± 0	1 ± 0	1 ± 0
2 ± 1	**2** ± 0.4	1 ± 0	1 ± 0
1 ± 0	**2** ± 0.3	**3** ± 0.4	1 ± 0
1 ± 0	1 ± 0	**5** ± 0.7	**12** ± 2

D-3P model

1	1	1
1	C_1	1
1	C_2	C_3

QIM best model (28%)

1 ± 0	1 ± 0	1 ± 0
1 ± 0	**3** ± 0.1	1 ± 0
1 ± 0	**6** ± 1	**23** ± 11

B-5P model

C_1	1	1
B_1	C_2	1
1	B_2	C_3

QIM model (21%)

3 ± 1	1 ± 0	1 ± 0
2 ± 0.4	**4** ± 0.1	1 ± 0
1 ± 0	**9** ± 2	**35** ± 17

Figure 1. A: QInfomating user-friendly interface. B: Best-fitting models in the current analysis and previous work (see text).

The best model by Lau et al. had a weight of 99%, so it is almost exact to the multi-model inference parameter matrix. In the case of QInfoMating, the best model (D-3P) had only 28% of the weight and other models such as B-5P (21%), D-5P (21%), D-4P (18%) and D-6P (8%), are necessary to capture more than 95% of the weight. The pattern obtained using the Bayesian Information Criterion (BIC) was similar to that obtained using the AIC criterion but with greater weight concentrated in the D-3P model (Table 3).

However, the structure of all these models is very similar: large males (rightmost columns) choose females larger (closest to the bottom rows) than themselves, especially for the largest classes (B_2 and B_3 parameters in the bias models, or c_2, c_3 in the D-3P model) .

The pattern behind the QInfoMating models aligns with the findings reported by Lau et al. [6] for the Multiple-design and Wild data, as depicted in their Figure 4. However, there is a notable distinction: Lau et al. observed that the shorter class exhibited considerably lower mating ability. This difference may be attributed to the presence of their first class with a minimal number of individuals for both females and males, as documented in Table 1.

However, one might wonder why QInfoMating did not clearly favor the B-5P model, which is most akin to the B-7P model chosen in the original analysis by Lau et al [6]. The explanation lies in how QInfoMating operates when handling quantitative data. As explained before, the program selects models based on the statistical significance of various tests and have included double-effect models (as D-3P, D-4P and D-6P) among the set of models to be tested. These particular models were not considered in the original Lau et al. study.

To further elucidate, I excluded the double-effect models and repeated the analysis using exactly the same set of models as in Lau et al. Upon conducting this analysis, the best-fitting model becomes the B-5P, with a weight of 94%. The program's preference for the D-3P model, when included, stems from its ability to explain the data with two fewer parameters, thus aligning with the principle of parsimony. In cases where both models are closely weighted, the choice between them is influenced by the researcher's experience and the biological plausibility of the results. Nonetheless, it's worth noting

that the fundamental structure of these models remains quite similar: larger males tend to prefer females larger than themselves, whenever such a pairing is possible.

Consistent with the result of models, the statistical tests performed by QInfoMating detect assortative mating and sexual selection both in females and males (Table 4).

Discussion

Assortative mating based on size and other quantitative traits is common in many species, from gastropods [11] to humans [16]. The processes leading to assortative mating may not solely rely on similarity-based mechanisms and may not affect different phenotypic classes equally. This is where the power of model selection methodology becomes evident, as it doesn't require proposing a single ad hoc model. Instead, it allows best fit of data for a wide range of models, enabling researchers to study non-trivial mating patterns, some of which may not have been obvious before the experiment.

Conclusion

In conclusion, the application of information theory, model selection, and parameter estimation using multimodel inference are presented as powerful tools for the analysis of mating data, whether quantitative or categorical. QInfoMating is software designed to perform this type of analysis.

Availability and requirements

Project name: QInfomating

Project home page: https://acraaj.webs.uvigo.es/InfoMating/QInfomating.htm

Operating system(s): Windows, Linux, macOS

Programming language: C++, Python.

License: GNU GPL.

Declarations

- Data availability

  The datasets generated and/or analysed during the current study are available in the DRYAD repository, http://doi.org/10.5061/dryad.h214h8t

- Competing interests

  The authors declare that they have no competing interests

- Funding

  This was funded by Xunta de Galicia (ED431C 2024/22), Ministerio de Ciencia e Innovación (PID2022-137935NB-I00) and Centro singular de investigación de Galicia accreditation 2024-2027 (ED431G 2023/07) and “ERDF A way of making Europe”.

- Authors' contributions

  ACR is the sole author of this work.

- Acknowledgments

  Dedicated to my sister Paci. I thank E. Rolán-Alvarez for comments on the manuscript.

References

1. Henshaw JM, Kahn AT, Fritzsche K. A rigorous comparison of sexual selection indexes via simulations of diverse mating systems. Proc Natl Acad Sci. 2016;113:E300–8.

2. Carvajal-Rodríguez A. Unifying quantification methods for sexual selection and assortative mating using information theory. Theor Popul Biol. 2024;158:206–15.

3. Carvajal-Rodríguez A. Non-random mating and information theory. Theor Popul Biol. 2018;120:103–13.

4. Carvajal-Rodríguez A. Multi-model inference of non-random mating from an information theoretic approach. Theor Popul Biol. 2020;131:38–53.

5. Estévez D, Kozminsky E, Carvajal-Rodríguez A, Caballero A, Faria R, Galindo J, et al. Mate Choice Contributes to the Maintenance of Shell Color Polymorphism in a Marine Snail via Frequency-Dependent Sexual Selection. Front Mar Sci. 2020;7.

6. Lau SLY, Williams GA, Carvajal-Rodríguez A, Rolán-Alvarez E. An integrated approach to infer the mechanisms of mate choice for size. Anim Behav. 2021;175:33–43.

7. Carvajal-Rodríguez A, Rolan-Alvarez E. JMATING: a software for the analysis of sexual selection and sexual isolation effects from mating frequency data. BMC Evol Biol. 2006;6:40.

8. Evren A, Tuna E. On some properties of goodness of fit measures based on statistical entropy. Int J Res Rev Appl Sci. 2012;13:192–205.

9. Kullback S. Information Theory and Statistics. New edition. Mineola, N.Y: Dover Publications; 1997.

10. Erlandsson J, Rolán-Alvarez E. Sexual selection and assortative mating by size and their roles in the maintenance of a polymorphism in Swedish Littorina saxatilis populations. Hydrobiologia. 1998;378:59–69.

11. Ng TPT, Rolán-Alvarez E, Dahlén SS, Davies MS, Estévez D, Stafford R, et al. The causal relationship between sexual selection and sexual size dimorphism in marine gastropods. Anim Behav. 2019;148:53–62.

12. Caballero A. Quantitative Genetics. Cambridge University Press; 2020.

13. Lewontin RC, Prout T. Estimation of the Number of Different Classes in a Population. Biometrics. 1956;12:211–23.

14. Reid DG. The genus Echinolittorina Habe, 1956 (Gastropoda: Littorinidae) in the Indo-West Pacific Ocean. Zootaxa. 2007;1420:1–161.

15. Rolán-Alvarez E, Williams G, Ng T, Estévez-Barcia D, Davies M, Saltin S, et al. Data for: The causal relationship between sexual selection and sexual size dimorphism in marine gastropods. 2021;1.

16. Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, et al. Genetic evidence of assortative mating in humans. Nat Hum Behav. 2017;1:1–13.

Tables

Table 1. Mating table for the Wild data used in Lau et al [6]. The data is classified in four classes.

Premating males numbers (421)			
23	181	167	50
Premating female numbers (455)			
23	146	190	96
Matings (228, rows are females)			
0	2	0	0
3	30	14	1
2	42	45	9
1	8	43	28

Table 2. Mating table generated by QInfoMating for the Wild data in Lau et al (2021). The data is classified in three classes.

Premating males numbers (421)		
76	333	12
Premating female numbers (455)		
57	370	28
Matings (228, rows are females)		
4	6	0
18	163	2
1	30	4

Table 3. Models that accumulate up to 95% of the weight. The parameter notation corresponds to the B-7P model in Figure 1B where C_i represent the ith cell on main diagonal and B_i on the subdiagonal.

Analysis	Models (>95%)	Weight	Parameter estimation						
			C_1	C_2	C_3	C_4	B_1	B_2	B_3
Lau et al 2021	B-7P	99%	0	2	3	12	2	2	5
QInfoMating-AIC	D-3P	28%	1	3	22	-	1	6	-
QInfoMating-AIC	B-5P	21%	3	4	35	-	2	9	-
QInfoMating-AIC	D-5P	21%	3	4	35	-	2	9	-
QInfoMating-AIC	D-4P	18%	2	3	24	-	1	6	-
QInfoMating-AIC	D-6P	8%	3	4	36	-	2	10	-
QInfoMating-BIC	D-3P	85%	1	3	23	-	1	6	-
QInfoMating-BIC	D-4P	10%	2	3	24	-	1	6	-
QInfoMating-BIC	B-5P	2%	3	4	35	-	2	9	-
QInfoMating-BIC	D-5P	2%	3	4	35	-	2	9	-

Table 4. QInfoMating tests.

Test type	Statistic	H_0	*p*-value
Quantitative	$QJ_{PSI} = 0.49$	no assortative mating	$<10^{-6}$
Quantitative	$QJ_{PS1} = 0.24$	no female sexual selection	$<10^{-6}$
Quantitative	$QJ_{PS2} = 0.07$	no male sexual selection	0.0004
Discrete	$J_{PSI} = 0.08$	no assortative mating	0.0009
Discrete	$J_{PS1} = 0.17$	no female sexual selection	$<10^{-6}$
Discrete	$J_{PS2} = 0.06$	no male sexual selection	0.002