Is the Genotype-Phenotype Map Modular?: A Statistical Approach Using Mouse Quantitative Trait Loci Data

Jason G. Mezey,*1 James M. Cheverud² and Günter P. Wagner*

*Department of Ecology and Evolutionary Biology, Center for Computational Ecology, Yale University, New Haven, Connecticut 06520-8106 and ²Department of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, Missouri 63110

Manuscript received November 5, 1999
Accepted for publication May 24, 2000

ABSTRACT

Various theories about the evolution of complex characters make predictions about the statistical distribution of genetic effects on phenotypic characters, also called the genotype-phenotype map. With the advent of QTL technology, data about these distributions are becoming available. In this article, we propose simple tests for the prediction that functionally integrated characters have a modular genotype-phenotype map. The test is applied to QTL data on the mouse mandible. The results provide statistical support for the notion that the ascending ramus region of the mandible is modularized. A data set comprising the effects of QTL on a more extensive portion of the phenotype is required to determine if the alveolar region of the mandible is also modularized.

IT has been hypothesized that characters that collectively serve a common functional role should be genetically integrated and relatively independent from the rest of the body (Olson and Miller 1958; Riedl 1978; Cheverud 1982; Bonner 1988; Wagner and Altenberg 1996). The genetic organization of a complex of characters thus is expected to reflect functional relationships. This is also called the hypothesis of the “imitatory epigenotype” because it is predicted that the pattern of developmental constraints “imitates” the pattern of functional constraints (Riedl 1978). Closely associated with the hypothesis of an imitatory epigenotype is the concept of a module (Raff 1996; Wagner and Altenberg 1996). A module is defined as a complex of phenotypic traits that is both tightly integrated by pleiotropic effects and relatively independent of the rest of the phenotype (Wagner 1996). So defined, a module is recognized by the statistical distribution of mutation effects among phenotypic traits. The imitatory epigenotype hypothesis thus predicts that functional units of the phenotype should be genetically modular.

The study of Cheverud et al. (1997) addresses the issue of modularity with a study of the mouse mandible (Figure 1). A quantitative trait locus (QTL) study was performed to determine the effects of alleles at individual loci. For the QTL with more than two effects (21 in total), the authors determined that only a minority of 23% had effects on both the ascending ramus (traits 1–11 and 21) and the alveolar region (traits 10 and 12–20) of the mandible. Furthermore, most QTL with effects on more than two traits had effects significantly focused on either the ascending ramus (50%) or the alveolar region (27%). These results were interpreted as consistent with the theory that selection tends to produce modular structure in functionally distinct units. However, without analyzing these findings in a statistical framework for all the loci simultaneously, it is difficult to determine whether 27 and 50% of QTL biased to the ascending ramus or the alveolar region are significantly different than expected under a random model.

In this study we propose an additional method for assessing the modularity of a portion of the phenotype using QTL data. We suggest two statistics that measure the two aspects of modularity: integration by pleiotropic effects and parcelation, i.e., fewer than expected pleiotropic couplings with the rest of the phenotype (Wagner and Altenberg 1996). These statistics are used to test whether units of the phenotype are more modular than expected by chance. In this article, we apply the test to the same QTL reported in Cheverud et al. (1997) and find that the ascending ramus is more modular than expected by chance.

THE UTILITY OF QTL DATA FOR EXPLORING THE GENOTYPE-PHENOTYPE MAP

Hypotheses about the structure of the genotype-phenotype map are statements about the statistical distribution of mutation effects among phenotypic traits. Ideally, data to test these hypotheses should directly sample the distribution of the effects of individual mutations. Data on the distribution of individual mutation effects, however, is difficult to produce. QTL data provide a reasonable alternative for testing ideas about the struc-
manner are interpreted as being closely linked to loci with an effect on the given traits. Assuming a proper density of molecular markers along the chromosomes, the marker locus technique allows identification of a large number of QTL that may be located anywhere in the genome. The technique may be used to identify the set of genes causing heritable differences between lines and the individual pleiotropic effects of each gene (Cheverud et al. 1997).

The marker locus technique identifies loci where different alleles have become fixed in divergent lines. Whether a particular allele becomes fixed in a line depends on two probabilities: (1) the probability that a mutation with this effect will occur, and (2) the probability of fixation given that the mutation has occurred. The first probability is relevant for estimating the distribution of mutation effects. It is this statistical distribution that is predicted to be modular. The second probability is determined by the history of drift and selection effects in the study populations. Of the mutations that were segregating in the original base population and the mutations that occurred during selection, some are more likely to be fixed than others. For example, if a line used in a QTL study is produced by selecting for larger character values, mutations with large positive effects on the character are more likely to be fixed. The result is that marker locus QTL data are a "filtered" sample of the mutational distribution.

For the purposes of this study, we are interested in marker locus studies where multiple traits were measured and a large number of QTL were identified. In this study, we reanalyze the data set of Cheverud et al. (1997) that identified QTL with effects on the mouse mandible. The inbred lines used in this study were produced from lines selected for large and small body weight, respectively. Hence, the lines were selected for alleles with large overall effect on body size. As explained below, we therefore think that this data set is biased against the modularity hypothesis and we consider the test results reported here as conservative.

STATISTICS FOR MODULARITY

Modularity of a set of traits can be defined by two attributes (Wagner and Altenberg 1996): (1) a higher than average level of integration by pleiotropic effects of genes among the traits of a set, and (2) a higher than average level of independence from other trait sets. The first of these conditions reflects the degree of integration within a module and the second reflects the degree of parcelation of the phenotype into distinct units. We assess the modularity of a set of traits $T_i$ that is a proper subset of the total number of traits observed $T$. Ideally, the number of traits in the subset is much smaller than the total set of observables. We define two statistics for assessing the modularity of $T_i$, one to measure integration within the set and one to measure
parcellation between trait sets. We call the total number of QTL that have effects on at least two traits $n$. Of these $n$ QTL, we call $m$ the number of QTL that affect at least one trait in $T_i, m < n$. Note that we only consider QTL with effects on at least two traits in $T$ when calculating these statistics.

**Integration statistic:** To measure integration of the set $T_i$ consisting of $k$ traits, we consider only the $m$ QTL with an effect on at least one trait in the set $T_i$. The integration statistic $I_{T_i}$ is based on the total number of traits in $T_i$ affected by the $m$ QTL compared to the maximum number of traits these QTL could affect in $T_i$, which is $km$:

$$I_{T_i} = \frac{\sum_{j=1}^{m} \text{(no. traits in } T_i \text{ affected by QTL } j) - m}{m(k - 1)}.$$  

The statistic is scaled to the unit interval $[0, 1]$. The extreme values $I_{T_i} = 1$ and $I_{T_i} = 0$ correspond to the maximum and minimum amounts of integration of the traits in $T_i$. For example, the most completely integrated case would be one where all $m$ QTL with at least one effect on $T_i$ have effects on all $k$ traits in $T_i$ such that the total number of effects is $km$. This corresponds to $I_{T_i} = 1$ (Figure 2A). At the other extreme, each QTL with at least one effect on $T_i$ affects only a single trait. In this case, each trait in $T_i$ changes independently as a result of an allele substitution. This corresponds to the lowest possible amount of integration such that $I_{T_i} = 0$ (Figure 2A).

**Parcelation statistic:** For measuring parcelation, we consider all $n$ QTL in the sample with effects on more than one trait. Parcelation is defined as a relative lack of pleiotropic effects between two sets of traits. Parcelation is thus a relation between any two nonoverlapping sets of traits $T_1$ and $T_2$. These two sets can be a subset $T_j$ of all traits $T$ and its complement, $i.e.,$ all the traits in $T$ that are not in $T_j$. Pearson’s $\chi^2$ is used to compare the observed number of traits affected by each QTL in $T_1$ and $T_2$ to the expected number based on the marginal distribution of effects among all QTL and traits. For each QTL $j$, the observed number of traits affected in $T_1$ and $T_2$ is symbolized by $O_{j1}$ and $O_{j2}$, and the expected number of traits affected is calculated as $E_{j1} = p_1 S_j$ and $E_{j2} = p_2 S_j$, where $S_j$ is the total number of effects of QTL $j$, and $p_1$ and $p_2$ are the relative frequencies with which a trait in $T_1$ or $T_2$ is affected by any QTL in the total sample,

$$p_1 = \frac{\sum_{j=1}^{n} O_{j1}}{\sum_{j=1}^{n} (O_{j1} + O_{j2})} \quad \text{and} \quad p_2 = 1 - p_1.$$

The parcelation statistic $M_k$ therefore reads

$$M_k = \sum_{j=1}^{n} \frac{(O_{j1} - E_{j1})^2}{E_{j1}}.$$

Intuitively, this statistic measures the degree to which QTL preferentially affect traits in either $T_1$ or $T_2$. The null hypothesis is that the effects of a QTL are randomly distributed among the traits of the two trait sets. Rejecting the null hypothesis means that QTL effects tend
to be clustered within each trait set. In genetic terms, this means that the QTL are “character specific,” i.e., their effects tend to be limited to a set of traits. This was tested in Cheverud et al. (1997) for each individual QTL. The present statistic assesses this association between QTL and trait sets for the whole data set. The larger the value of this statistic, the less likely it is that a mutation will have an effect on both trait sets \( T_i \) and \( T_j \) (Figure 2, B and C).

**Application and test distribution:** To illustrate how these statistics are utilized, consider a study where \( n \) QTL have been identified that have effects on at least two of \( N \) observed phenotypic traits. Let us assume that \( k \) of these traits describe a complex \( T_i \) that performs some common function. The information about the functional significance of \( T_i \) must be available from some independent source, such as a study of the functional morphology of the traits. We want to test whether \( T_i \) is more modular than expected for a random set of traits. To do this, we first determine the values of the statistics for \( T_i \) and then the cumulative mass function (cmf) of \( M_i \) and \( M_p \) with respect to all the possible subsets of \( k \) traits taken from the total \( N \) traits. If both the values of \( M_i \) and \( M_p \) are greater than the critical value (using \( \alpha = 0.05 \) for this study), the null hypothesis is rejected; i.e., the modularity of the set is greater than expected for a randomly chosen set of traits.

The “shape” of the cmf for the statistics \( M_i \) and \( M_p \) will depend on the distribution of pleiotropic effects among the QTL. Optimally, the value of \( M_i \) and \( M_p \) for all possible sets of \( k \) traits should be determined to calculate the cmfs exactly. This is the procedure adopted in our study. However, if the values of the \( E_y \) are generally \( >5 \) and if no or only a few QTL have a number of effects \( >k \) or \( N - k \), then the sampling distribution \( M_p \) will approximately follow a \( \chi^2 \) distribution with \( 2n \) d.f.

We do not in general expect the marginal distributions of the statistics \( M_i \) and \( M_p \) to be independent. If they are not independent, the probability of rejecting the null hypothesis for both statistics for a random set of \( k \) traits is somewhere between 0.0025 (for independent \( M_i \) and \( M_p \)) and 0.05 (for a correlation of 1.0). Regardless of the correlation structure, we still consider a functional unit to be significantly modular if we can reject the null hypothesis for both statistics. It is expected that the correlation of the statistics will not be useful for comparing the power of the test for different data sets as the distribution associated with an alternative hypothesis (how traits of the unit \( T_i \) are assigned) will not necessarily be comparable between data sets.

It should be noted that for cases where we are assessing the integration and parcelation of more than one unit of a single data set, the units should not have many traits in common. If two units have many traits in common, we would expect similar results (rejection or nonrejection) for both units. We must therefore be cautious when assessing hierarchical and overlapping sets of traits for their degree of integration or parcelation. In this case, the two units of the mandible have only a single trait in common.

**RESULTS**

For this study, we assessed the modularity of functional sets of the mandible of the mouse *Mus musculus*. A QTL study performed by Cheverud et al. (1997) discovered 41 QTL, of which 37 QTL affect at least 2 of the 21 traits of the mandible (Figure 1). A summary of the amount of pleiotropy in this data set (the frequency of QTL affecting \( z \) traits) is provided in Figure 3. The average number of traits affected by a QTL is 4.0, and the maximum is 13 traits. In this study, we assessed two functional sets of traits defined in Cheverud (2000), the ascending ramus (traits 1–11 and 21; \( k = 12 \)) and the alveolar region (traits 10 and 12–20; \( k = 10 \); Figure 2). All the major mandibular muscles attach to the ascending ramus. The shape of the ramus is critical for proper motion of the chewing and biting apparatus and therefore is considered to have been under strong selection pressure (Atchley and Hall 1991). The alveolar region is the location of tooth attachment. The form of this region is critical for proper positioning of the teeth (Atchley et al. 1992). The results discussed below are summarized in Table 1.

**Ascending ramus:** The ascending ramus is described by \( k = 12 \) of the 21 traits in this study. These traits are affected by \( m = 32 \) QTL of the 37 that have effects on at least 2 traits. The maximum number of traits these QTL could affect is \( 12 \times 32 = 384 \). The observed number of effects on ascending ramus traits is 98 and thus the integration statistic is \( M_i = 0.188 \). The cmf of the integration statistics for \( k = 12 \) yields a \( P \) value of 0.0102, which is significant at the 2% level. This value reflects the probability that this or greater levels of integration of traits for their degree of integration or parcelation. In this case, the two units of the mandible have only a single trait in common.
of traits. As such, this test does not address which evolutionary forces may have caused the nonrandom pattern. If, however, the test is performed on a set of traits that a priori have been classified as describing a functional unit, it is in fact an indirect test of the hypothesis that natural selection has shaped the genotype-phenotype map to modularize functionally integrated sets of characteristics. A rejection of the null hypothesis is considered as evidence in support of the hypothesis of the imitative epigenotype (RIEDL 1978). In drawing this conclusion, it is of critical importance that the classification of the traits into functional units is done on the basis of independent evidence (see SCHWENK 2000 for a review of the concept of a functional unit). This is the case in this study, where the ascending ramus and alveolar region of the mandible have been identified as functional units independent of the genetic structure (ATCHLEY and HALL 1991; ATCHLEY et al. 1992).

For the data set of CHEVERUD et al. (1997), the ascending ramus represents a well-modularized portion of the phenotype. For the alveolar region, rejection of the parcelation null hypothesis indicates that mutation effects are significantly parceled with respect to this region. Although the integration null hypothesis could not be rejected for the alveolar region, the P value is only 0.0968 (Table 1). Given the significant test results for the ascending ramus, the results for the alveolar region (significant result for parcelation and nonsignificant result for integration) are somewhat expected. Since the ascending ramus and alveolar region together comprise all 21 traits of the mandible, the parcelation statistic is measuring the degree to which QTL affect either the ramus or the alveolar region. Hence, if we observe a significant result for one region we are likely to observe a significant result for the other. Similarly, since the ascending ramus is highly integrated, the alveolar region would need to have a level of integration close to the ramus to produce a significant result as well. That is, both units may be integrated but the effect will only be detected if the integration of the units is comparable. A better test would consider more traits such that together the ascending ramus and alveolar regions would contain far fewer than the total number of traits. If this is the case, different levels of integration may still produce a significant result in the integration test and we might expect different results for the parcelation test. To determine if both of these units reflect a modular pattern, we would therefore need a QTL data set comprising more traits. Our study does at least confirm that the ascending ramus is a genetically modular unit of the phenotype (CHEVERUD et al. 1997).

A limitation of our tests is that they only take into account whether a QTL has a significant effect on a trait or not. Ideally one would like to take into account the magnitude of the effects to assess the degree of integration and/or parcelation. The problem, though, with including the magnitude of the effects is that most of the effects are small and thus the confidence intervals are relatively large. This implies that there is little actual information added to the analysis if the measured effect is taken into account other than the fact that the effect is greater than zero.

As explained previously, every QTL data set repre-

<table>
<thead>
<tr>
<th>Region</th>
<th>Value of $M_1$ (P value)</th>
<th>Value of $M_0$ (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending ramus</td>
<td>0.188 (0.0102)</td>
<td>77.6501 (0.0003)</td>
</tr>
<tr>
<td>Alveolar region</td>
<td>0.171 (0.0968, NS)</td>
<td>68.6876 (&lt;0.0001)</td>
</tr>
</tbody>
</table>

NS, not significant.
sents a filtered sample of mutations that may occur. That is, a QTL data set not only reflects the probability of a mutation occurring but also the fixation probability of a mutation during the preparation of the lines used in the QTL study. Since different QTL study conditions may produce different probabilities of fixation, it is possible that the tests may produce a significant result for one set of lines but may not with another set of lines. Consider, for instance, a functional complex that is modular. A group of mutations that occur in this system are likely to reflect this modular structure. However, a QTL study protocol may end up fixing a filtered sample of these mutations that no longer have a modular distribution. Because QTL study conditions could also have the opposite effect, making a nonmodular unit appear modular, it is important to test data from other QTL studies using different inbred lines before drawing a final conclusion concerning the modularity of a structure. Since multiple QTL data sets do not presently exist for the mouse mandible, we can only interpret the present results as consistent with the hypothesis that the ascending ramus is a modular unit.

The genetic structure of the mouse mandible has been studied in the past with quantitative genetic methods (Bailey 1956; Atchley et al. 1985, 1992). Bailey (1956) presented a principle component analysis of the genetic correlations matrix and found a general size factor and one factor with contrasting loadings on alveolar region and ascending ramus measurements. This result is consistent with our finding that these two parts of the mandible are modular. Bailey's result, however, may in part be a measurement artifact since all the measurements in his study have one landmark in common, located at the junction between the ascending ramus and the corpus of the mandible. This can cause spurious correlations proportional to the angle among the distances measured. Atchley et al. (1985) presented a cluster analysis of genetic correlations among mandible traits, which showed two clusters of ascending ramus and alveolar measurements in addition to some independently varying traits. A reanalysis of the same data in Atchley et al. (1992) did not reveal any clear pattern of modularity (see Figure 9 in Atchley et al. 1992).

The comparison of genetic correlations and QTL data on the genetic architecture of characters is complicated by two factors. At the one hand, the genetic correlations reflect not only the pattern of mutational effects but also the frequencies of segregating alleles in the population in which the genetic correlations were estimated. Genetic correlations are as much a property of the population composition as they are a property of the genetic architecture of the traits. Furthermore, a low genetic correlation can indicate one of two things: it could be caused by a relative lack of pleiotropic effects among the traits or a pattern of positive and negative pleiotropic effects that cancel each other. There is no way to distinguish between these two scenarios by analyzing genetic correlations. Our results and that reported in Cheverud et al. (1997) show that the independent variation between the two major regions of the mandible is due to a relative lack of pleiotropic effects rather than due to antagonistic pleiotropic effects.

It should be noted that although the results from the QTL analysis are consistent with the hypothesis that natural selection may result in modular structures, results from this test alone are not sufficient to actually infer the action of natural selection in creating a modular pattern. To more directly investigate the possible role of natural selection in causing the association between functional and genetic integration, it is necessary to analyze data on the evolution of the genotype-phenotype map and the evolution of character function. However, with analyses of the sort described in this study, we can now begin to examine the architecture of genotype-phenotype maps and provide a foundation for such studies.

The authors thank Junhyong Kim and David Houle for critically reading the manuscript. We further thank Homayoun Bagheri, Chi-Hua Chiu, Ashley Carter, Christian Pazmandi, Scott Rifkin, Max Shpak, and Terri Williams for stimulating discussions on the topic of this article. The support by National Science Foundation grants IBN-9905403, DEB-9419992, and DEB-9726453 is gratefully acknowledged.

LITERATURE CITED


Communicating editor: Z-B. Zeng