# LIMITS TO KNOWLEDGE IN POPULATION GENETICS

Michael T. Clegg

Department of Botany & Plant Sciences

University of California

Riverside, California 92521

This contribution appeared in Evolutionary Biology, Vol 32 pp35- 51. Volume 32 of Evolution Biology contains 13 articles by prominent evolutionary geneticists on the theme of Limits fo Knowledge in Evolutionary Genetics.

FAX (909) 787-4190

E-MAIL: Clegg@ucrac1.ucr.edu

CONTENTS

# INTRODUCTION

The modern science of biology traces, in large part, to the publication of the "Origin of Species" by Charles Darwin. By invoking natural selection as the causal agent of biological change, Darwin provided a mechanistic explanation for the elaboration of biological diversity that focused on the experimental manipulation of observable processes. In one stroke biology was transformed from a purely descriptive enterprise concerned with cataloging the wonders of biological diversity discovered during the age of exploration to an experimental science. The second major impact of the publication of the "Origin of Species" was to focus late 19th century biology on the great unsolved puzzle of hereditary transmission. It is, of course, well known that the solution to this great puzzle was published within a decade of the Origin, but lay unappreciated for more than thirty years. It is also well known that the first two decades of the 20th century were consumed by a debate about whether the selection of minute continuous variations (as postulated by Darwin) could be consistent with the particulate system of inheritance described by Mendel.

Theoretical population genetics originated as a mathematical consistency argument to establish that natural selection could be effective under the Mendelian scheme of inheritance (Provine, 1971). This program was immensely successful because it showed that a particulate system of inheritance was in fact essential for the conservation of genetic variance; a fundamental prerequisite for natural selection to be effective in driving evolutionary change (Fisher, 1930).

Population genetics focused on the gene as the unit of evolutionary change. Assuming particular mating schemes that determine the distribution of genotypic frequencies in each new generation and assuming various parameters describing mutation, migration, selection and the stochastic sampling of genes transmitted to each new generation (genetic drift), equations were derived to predict the trajectory of gene frequency change. This theoretical construct continues to be elaborated to the present day. It is important to appreciate, as emphasized by Wright (1931), that the models of population genetics are at best highly simplified descriptions of the complexity of actual population dynamics. It is simply not possible to incorporate the complexities of variable selection driven by temporal and spatial environmental change together with drift, migration, and demographic structure into a comprehensive model that is susceptible to analysis. Yet, evolution is actually determined by the continuous operation of these and other factors. Thus, while the conclusions of population genetic models are mathematically true, they may not always contribute to empirical knowledge.

Experimental population genetics began in earnest in the 1930s. The initial goals were to detect selection and to estimate the intensity and pattern of selection. Other goals were to measure migration rates, to estimate effective population sizes and to determine the overall extent of genetic variation within populations. In other words, the goals were to determine whether levels of genetic variation were sufficient for effective natural selection and to estimate the parameters that indexed the models of theoretical population genetics. The primary issue of concern was to determine the importance and strength of selection to provide empirical confirmation for theoretical calculations. Much of this program was embodied in the monumental “Genetics of Natural Populations” series initiated by Dobzhansky and his various collaborators (Dobzhansky, 1981).

This early work was largely experimental in the sense that populations were subject to manipulation and the quantities of interest were measured. Over the years it became clear that some of the parameters of interest could not be estimated either because experimental error is too large or because parameters like effective population size and selection are hopelessly confounded (Lewontin, 1974). In more recent years, the focus has moved from manipulative experiments to historical inference. The interest in historical inference has been stimulated by molecular data that permit direct comparisons of gene sequences among species. These data are then used to infer phylogenies or to infer the evolutionary history of particular genes and gene families. There are strict limits to historical inference. It is not possible to "know" that the particular history estimated is the true history, although we may be able to say with a given probability that the true history lies within some ensemble of histories.

To summarize, mathematical population genetics is over simplified. The theory defines the conditions for evolutionary advance, but some of the most important parameters can not be estimated with any precision and are at best knowable in only the crudest empirical sense. Moreover, the goal of historical inference is only approximately attainable and yet contemporary molecular systematics appears to assume that complete knowledge is attainable. It is also important to realize that the limits to knowledge in population genetics, and hence also the limits to knowledge in the allied sciences of molecular evolution and conservation biology, are not obvious. Workers in these fields have only come to a partial appreciation for the limits of knowledge as they have explored the empirical content of population genetic theory over a period of more than sixty years.

**A classification of limits to knowledge in population genetics**

To facilitate discussion of the limits of knowledge it is useful to consider a partial classification of the reasons why the answers to a proposition are outside the bounds of scientific knowledge. Such a classification would include the following points:

1. Knowable in principle, but not with present technology:
2. Knowable in principle, but the experimental error overwhelms any signal:
3. Knowable in principle, but the theoretical framework for interpretation is positively misleading:
4. Unknowable because parameters are confounded:
5. Unknowable because the stochastic process that governed the history has too great a variance:
6. Unknowable because the complexity of the biological system precludes deterministic prediction:

Knowledge is often approximate and contingent on special conditions obtaining. It is rarely complete and immutable. The exploration of new knowledge occurs on a boundrary that defines the currently known and unknown. This boundrary is rarely sharp and well defined. In research we probe the boundary by devising new methods and new theoretical constructs that circumscribe problems and render them tractable. The objective of this chapter is to explore two contemporary issues in population genetics as a means of probing this boundary region. The first issue arises from the application of coalescence theory to the problem of inferring the historical processes that lead to a particular gene genealogy (realization). The second issue we explore is the problem of genetic / ontogenetic complexity. Specifically I will explore the way genes interact in determining phenotype via their metabolic and developmental context.

## THE ANALYSIS OF GENE GENEALOGIES

Much of the classical theory of population genetics was framed in terms of the transmission of genes that are identical-by-descent. The conceptual framework allows genes to change state through mutation, but their path of descent is assumed to be known. It is then relatively easy to derive transition equations that incorporate sampling effects (genetic drift), mutation and migration for identical-by-descent systems (see e.g. Crow and Kimura, 1970). It is much more difficult to incorporate selection because selection causes the systems of equations to be nonlinear and limits the scope of analysis. Until recently the empirical technology did not permit geneticists to infer patterns of descent for genes sampled from real populations. The best that could be inferred was that two genes are identical-in-state, but identity-in-state does not imply identity-by-descent.

To provide a concrete example, we consider the isozyme technology of the 1960s and 70s: a particular isozyme state is determined by the net charge and conformation of the protein molecule. (The isozyme state is defined as the location of a particular enzymatic protein on a gel medium, where the proteins have been separated by electrophoresis.) Net charge is in turn, the result of both the pH of the medium and the distribution of acidic and basic amino acids comprising the protein molecule. A mutational change that alters a single amino acid may induce a location on a gel (state) that is identical to other very different amino acid configurations (Veuille and King, 1995). Hence, comigration of a band (identity in location on a gel) does not imply genetic similarity and does not permit the estimation of identity-by-descent relations. In this case the identical-by-descent relationships are unknowable in the absence of detailed information on the breeding history of individuals in the population. The isozyme example illustrates the situation where the identity status may be knowable in principle, but it is not knowable based on the isozyme technology alone (point #1 above).

With today's technologies for rapid DNA sequencing identity-by-descent can be estimated with considerable accuracy, linking the objects of theory more closely to the quantities that can be observed in the empirical world. This follows because

DNA sequence data measure directly the mutational similarity between copies of a gene, hence patterns of descent (genealogies) can be estimated from the joint distribution of nucleotide differences between a set of homologous sequences (sequences that have descended from a common ancestral sequence). Thus a new technology has rendered the unknowable at least approximately knowable. An interesting further point about the technology of rapid DNA sequencing is that it yields the most elemental of all genetic data. Because these data are the most elemental (i.e., the base sequence of a gene), no further technology can be expected to provide enhanced genetic resolution. (Of course, there is much room for improving the speed and throughput rate of DNA sequencing technology, so we may expect much larger data sets in the future that will reduce uncertainty by increasing statistical power.)

Why is the identical-by-descent situation important? It is important because a great deal of information is latent in the genealogical history of a sample of alleles. A fundamental theorem of coalescence theory shows that when the genealogy is the realization of a process that is strictly a drift / mutation process then the entire distribution of the sample history can be calculated (Hudson, 1990). More generally however, the structure of a genealogy (estimated internode lengths, branch lengths, number of branches connecting the sample of DNA sequences) depends on the mutation rate ($\mu$), the effective population size ($N_e$), the pattern and intensity of selection, the history of recombination and the demography of the population ( Sawyer and Hartl, 1992; Hudson, 1993; Nee et al., 1995; Holmes et al., 1995). The distribution of the sample history can not be calculated for this more realistic scenario. Nevertheless, comparison of the estimated sample history for a set of DNA sequences to that calculated from elementary theory provides a means of integrating over long periods of evolutionary time so that episodes of relatively weak selection can be detected (i.e. $s<0.01$). What can not be measured is the exact pattern of selection, recombination and demography that actually produced the observed realization. To summarize, the analysis of gene genealogies exploits both large amounts of nucleotide sequence data and a well defined null hypothesis of no selection (s=0), but the analysis can not provide a precise (or even very approximate) reconstruction of the actual processes that produced a given history.

Thus, the invention of a new technology appears to have brought theoretical and empirical population genetics into much closer alignment and it has therefore moved some categories of questions from the unknowable to at least the crudely knowable. However, there are strict limits to the inferences that can be supported from these kinds of data. The limits fall into categories 2, 4 and 5 above and before elaborating on these limits to knowledge it is useful to review in more detail the conceptual tools available for the analysis of gene genealogies.

**Inferences based on coalescence theory**

Many different tests of gene sequence samples have been proposed and taken together these tests provide some basis for inferring the relative importance of the various evolutionary processes that have affected genetic variation. Under the assumption of a strict drift / mutation process (referred to below as “neutrality”), the time intervals between coalescence events (nodes of a genealogy), measured in mutational change, is dependent on $4N_e\mu = \theta$. Thus the estimation of $\theta$ is central to testing the null hypothesis of neutrality. Assuming neutrality is accepted, and assuming that the neutral mutation rate is known, the effective population size of the species can then be estimated. [The efficient estimation of $\theta$ is a nontrivial problem because the intercorrelation of sequence change has to be accounted for by estimating $\theta$ and the genealogy together. Moreover, the possibility of recombination among homologous sequences must also be built into the estimation model (Kuhner et al., 1995).]

Several statistics that are simple functions of $\theta$ are used as the basis for statistical tests of a neutral genealogy. Among these statistics are S, the number of segregating sites where $E(S) = a_n \theta$, k the average number of pairwise nucleotide differences where $E(k) = \theta$ , and $\eta_s$, the number of singletons in the sample, where $E(\eta_s) = [n/(n-1)]\theta$. (Singletons are mutations that occur in only one sampled allele.) And, where $a_n = \bullet_{i=1,n}(1/a_i)$, n is the number of observations in the sample and E indicates the operator for mathematical expectation. Each of these expressions is an unbiased estimator of $\theta$ (Watterson, 1975).

A test statistic introduced by Tajima (1989) is based on the difference between k and $S/a_n$ and test statistics introduced by Fu and Li (1993) are based on the difference of $S/a_n$ and $\eta_s$ [(n-1/n] and on the difference of k and $\eta_s$ [(n-1)/n]. Assuming neutrality, each of these test statistics has an expected value of approximately zero. However, the exact distribution of each test statistic is unknown. Critical values for statistical tests were suggested by Tajima (1989) based on assuming a beta distribution and by Fu and Li (1993) based on simulated data. As a consequence, the actual bases for statistical analysis are very approximate.

Simonsen et al. (1995) have carried out extensive studies of the power of these, and several analogous test statistics, under the specific alternatives of (a) a selective sweep, (b) a population bottleneck and, (c) population subdivision. The overall conclusion of this study was that Tajima's test is usually the most powerful test, however, the power to reject the neutral null hypothesis is limited for samples of fewer than 50 sequences of a gene, and when the sample includes ten or fewer observations the power is virtually nil. In the case of a selective sweep Simonsen et al. (1995) note that there is a specific window of time when the selective sweep can be detected and this corresponds to $T_s = 0.18$ to $T_s = 0.27$ (for $s = 10^{-4}$)

where $T_s$ is measured in $2N_e$ generations. (Thus for $N_e = 10^6$ the range is 360,000 to 540,000 generations.) This means that many episodes of selection could occur, but be undetectable because they fell outside the critical time range. In fact, for $T_s$ sufficiently large and $s=10^{-4}$ to $10^{-2}$, the test is actually less likely to accept the alternative hypothesis of selection (even though selection is present) than the null hypothesis of no selection! The final important conclusion of the work of Simonsen et al. (1995) is that more statistical power is almost always obtained by sampling more individuals rather than more nucleotide sites per individual. In light of these facts, it is not surprising that the null hypothesis (s=0) is rarely rejected for gene sequences sampled from within species because few studies approach sample sizes of 50 individuals. In this case, knowledge is limited by statistical power but the issue of power can in principle be overcome with sufficiently large sample sizes (# 2 in the above classification). Knowledge is also limited by the fact that for certain regions of the trajectory of gene frequency change the hypotheses of selection versus neutrality are not distinguishable because the two processes are virtually identical (corresponding the # 5 above). Finally, estimates of $N_e$ are always confounded when selection is present because selection amounts to a restriction of effective population size (# 4 above). There are other potential tests for selection , but all suffer from similar defects (Clegg, 1997).

**Some empircal data**

Cummings and Clegg (1998) used the brute force approach of sampling a large number of genes to address the power limitation. This work was based on a sample of 45 sequences from the alcohol dehydrogenase 1 (*adh* 1) locus of wild barley (*Hordeum vulgare* ssp. *spontaneum*). The various test statistics described above showed significant deviations from the null hypothesis of neutrality, so we may conclude that the strictly neutral model can be rejected. However, the deviation from neutrality appeared to be accounted for by an excess of singleton amino acid replacement polymorphisms which may indicate better conformance to the nearly-neutral hypothesis of Ohta (1973, 1992) or the background selection hypothesis of Charlesworth et al. (1993) both of which posit that a flux of mildly deleterious mutations controls the distribution of polymorphism.

To further interpret these data it is important to recall that barley is a predominantly self-fertilizing plant with estimated rates of self-fertilization in the neighborhood of 99% (Brown et al., 1978; Clegg et al., 1978). Because self-fertilization implies high levels of homozygosis, we expect the equilibrium frequencies of deleterious mutations to be approximated by $\mu/s$, so it is surprising to find such a high frequency of replacement mutations in the sample. Estimates of effective population size from synonymous and intron sites suggests $N_e$ is approximately $10^5$, so we would expect weak negative selection to be effective. It may be that population subdivision plays an important role in determining the distribution of replacement polymorphism. Under this assumption, selection within each small local population would be the factor determining the frequency of negatively selected variants. Some could then drift to moderate frequencies because local effective sizes may be small. Because the sample of 45 sequences was drawn from the whole geographical distribution of the species, it is not possible to test for the effects of population subdivision, however a gross examination of the geographical distribution of variation does not reveal obvious geographical patterns.

A straightforward effort to increase statistical power has not resolved the question, instead it has raised new questions about population substructure that require much larger sample sizes drawn from a geographically structured sample. Nor is it clear that such a sample would in fact resolve the question, because little is known of the history and demography of local populations.  So the answers are elusive and regress to other subsidary questions.

What lessons may be drawn from such experiments? It is like peeling an onion. There are layers of technical innovation, each promising to relate more closely to the objects of mathematical theory. At the same time the mathematical theory is elaborated layer after layer and yet the answers to the empirical questions that have motivated the science of population genetics remain elusive.

## ADAPTATIVE EVOLUTION IN COMPLEX BIOLOGICAL SYSTEMS

The fundamental theoretical quantity of population genetics is the gene, yet biological adaptations are not the isolated manifestation of individual genes. The target of selection is a phenotype which is determined by the interactions of many genes both in a temporal (developmental) and in a metabolic context. The methodological triumph of classical genetics was to breakout of this context dependency by studying extreme mutations that completely disrupted the function of a gene. By this means, it was possible to acquire an understanding of the "function" of a gene, but little was revealed about the subtle context dependent changes that are thought to be the substrate of adaptive evolutionary change. In contrast, theoretical population genetics simply assumed that selection produced adaptive change and then studied the consequences of selection on single genes or occasionally on pairs of genes. Both approaches beg the central empirical question of relating adaptive phenotypic change to its underlying genetic determinants. (This corresponds to point # 3 where knowledge is limited because the theoretical framework is misleading.)

### Chalcone synthase genes in morning glory

As a specific example of the complex interactions that arise in metabolism and development consider the flavonoid biosynthetic pathway of plants. This biochemical pathway is responsible for flower pigmentation and it also plays a major role in plant disease defense, lignin biosynthesis and in UV protection, so the mapping from genes to phenotypes is complex. Many genes are involved in determining flower color phenotypes, but those same genes also determine other distinct aspects of phenotype that are potentially adaptive. This complexity limits our ability to dissect the adaptive contribution of any particular mutation to a single aspect of phenotype (e.g. flower color, corresponding to point 6 of the above classification). To make this point concrete, Figure 1 illustrates the biochemical pathway that leads to flower pigmentation. Figure 1 also shows the various side branches that lead to other important compounds. A quick perusal of Figure 1 is sufficient to establish that many end products (phenotypes) may be dependent on the same enzymatic steps.

To facilitate our exploration of the limits to knowledge in this system, let us take as an example on going analyses of the evolution of the genes encoding one enzyme in this pathway within the common morning glory, *Ipomoea purpurea* . The first key step in flavonoid biosynthesis is the condensation of three molecules of malonyl CoA with one molecule of 4-coumaroyl CoA to form the 15 carbon naringenin chalcone molecule. The condensation is catalyzed by the enzyme chalcone synthase (CHS) which typically exists as a small multigene family in plant genomes (almost all plant genes appear to exist as multigene families).

**Gene duplication and divergence in evolution**

Durbin et al. (1995) characterized four genetic loci that encode *chs* genes in the morning glory (*chs* A, B, C, and PS). In addition, two other distantly related *chs* genes (D and E) have recently been described in the morning glory genome

(Fukada-Tanaka, et al., 1997). *chs* gene sequences have been determined from five *Ipomoea* species (Durbin et al. 1995; unpublished). Four of the *Ipomoea chs* genes (A, B, C and PS) appear to have duplicated approximately 25 million years ago, prior to the divergence of the *Ipomoea* species included in the sample. Based on genetic distances and nine characteristic amino acid substitutions, these four *chs* loci group into two related categories. Group 1 includes *chs* A and C while *chs* B and PS are contained in group 2. The amino acid substitutions observed in group 2 are characteristic of stilbene synthases (STS) and may indicate a functional divergence between duplicate *chs* genes. The independent evolution fo the STS function from CHS has occurred several times in flowering plant evolution (Tropf, 1994) and *chs* genes have evidently acquired other functions in flowering plant evolution as well (Durbin et al., 1999).

The *Ipomoea chs* genes appear to be evolving at a rapid rate and have a relatively low ratio of amino acid replacement to synonymous substitutions (about 5.6 synonymous to replacement changes, Glover et al, 1996; Durbin et al., 1999). Synonymous substitutions are presumed to be selectively neutral, however, replacement substitutions are more directly exposed to selection by virtue of their associated amino acid changes. Thus, the ratio of synonymous / replacement substitution serves as an index of the level of selective constraint operating upon sequences, with a lower ratio implying either reduced selective constraint or diversifying selection. At this stage we can only speculate on the potential adaptive bases for *chs* gene family expansion, but divergence in enzymatic function and divergence in developmental expression patterns (see below) appear to be common characteristics of *chs* gene family evolution.

The expression of the *chs* gene family members in *I. purpurea* is differentially regulated (Durbin et al., 1999). Four of the genes (*chs* A, B, C, and D) are known to be expressed based on the recovery of cDNA clones corresponding to each gene. However, *chs* A is found in the limb of the flower corolla while *chs* C is expressed in the throat of the corolla. *chs* mRNA is also found in mature anther tissue, although it is not established whether expression is specific to either pollen or anther tissue. Finally, *chs* expression occurs in the stem, pistil, cotyledon, and sepal tissues, but has not been observed in roots of young seedlings. These data are consistent with the multiple biological functions of the products of flavonoid biosynthesis. They also reveal considerable evolutionary divergence in developmental expression during the time since the elaboration of this gene family. We assume that diverging patterns of developmental expression are the result of adaptive responses to environmental conditions.

To summarize, the *chs* genes in *Ipomoea* have duplicated and diverged in function and they have diverged in specific expression patterns during *Ipomoea* evolution. All the evidence indicates that some of the *chs* genes have been subject to adaptive evolution, but the specific phenotypic effects associated with *chs* evolution remain obscure. We may have reached the limits of knowledge regarding the relationship between genetic change and adaptive phenotypic change in this system, but maybe not. It may be necessary to probe in more detail associated changes in the other genes of flavonoid biosynthesis that are functionally linked to *chs* in determining phenotype.

**Interlocus gene conversion / recombination**

Huttley et al. (1997) sampled and sequenced *chs* A genes from 14 lines that originated from collections in Mexico. (The center of origin of *I. purpurea* is in the highlands of central Mexico.) Four lines from Georgia and North Carolina were also sampled, where the plant is an introduced weedy species. No nucleotide sequence diversity was detected at *chs* A in the very limited sample from the US. The Mexican derived materials had much higher levels of nucleotide sequence diversity with eleven distinct haplotypes. Further examination of the *chs* A sequences reveals that the majority of variation resides in exons, as indicated by higher $\theta$ estimates for exon comparisons minus the flanking sequences and introns. Finally, a comparison between the *chs* A allele polymorphic sites indicates that at many of these sites one of the nucleotide states is present in non-*chs* A gene family members and these observations strongly suggest that both the high level of nucleotide diversity present in the *I. purpurea chs* A genealogy, and the relatively low ratio of synonymous to non-synonymous substitutions between *chs* A alleles, are probably derived from low to moderate rates of inter-locus recombination / gene conversion among the different *chs* gene family members (Huttley et al., 1997). This kind of observation is not unprecidented (see Chapter by Li in this volume), but it does represent a source of variation that is outside classical population genetic theory. In this sense our empirical knowledge has been expanded through the discovery of phonenoma that were not antiicpated by population genetics theory. But the question remains: can we determine the precise causal pathways through which adaptive evolutionary changes are mediated or is this beyond the methodological reach of contemporary biology?

**Genetics and population biology of white flower color phenotypes in morning glory**

Genes that modify the mating system by increasing rates of self-fertilization have been objects of theoretical interest because they are selected by virtue of their control of the system of genetic transmission (Fisher, 1941; Holsinger, 1996). Extensive experiments with the common morning glory over a period of many years have shown that insect pollinators discriminate against white flowers, which as a consequence, show higher rates of self-fertilization relative to fully pigmented flowers (Clegg and Epperson, 1987). A deeper analysis of selection on white flower phenotypes is complex and requires consideration of pollen discounting, maternal seed fertility, inbreeding depression among seed derived from self-fertilization relative to outcrossing and the effect of flower color frequency on the likelihood of self-fertilization (Holsinger, 1996). So far a completely comprehensive analysis has not been provided because the statistical power to detect modest selection ($s<0.05$) is limited when the scale of experiments is of a practical size.

The genetic situation is also complex because more than one locus may lead to a white flower color phenotype. For instance, a mutant phenotype that yields an albino flower (A/a locus, Epperson and Clegg, 1988) is characterized by the absence of the *chs* D transcript which suggests that *chs* D is uniquely responsible for the absence of anthocyanin pigments in the aa phenotype responsible for flower color (Durbin et al., 1999). However, the albino aa genotype is rare in populations (Epperson and Clegg, 1992) and a second locus (W/w; Epperson and Clegg, 1988) that yields a white flower color phenotype with pigmented floral rays is also descriminated against by pollinators. The W/w locus appears to be a regulatory locus based on genetic analysis (Ennos and Clegg, 1983) and on expression studies (unpublished data).

Selection on flower pigmentation by pollinators is about as simple a situation to analyze as possible, because the genetics and molecular biology are susceptible to analysis and the environmental agent of selection (bumblebees in this case) can be isolated and studied. Even in this case, however, the analysis is complicated by the fact that more than one gene is responsible for the phenopypes descriminated against in nature by the insect pollinators that effect plant reproduction. It is also complicated by considerations of experimental power. it is very probable that deeper analyses will reveal additional complexities in the causal pathway between selection and phenotypic determinaiton.

It is probable that the limitations to present knowledge are simply the consequence of inadequate experimental data and it seems reasonable to expect that further analyses will inform and enrichen our understanding of the causal pathways linking phenotypic diversity to adaptation. Nevertheless, the study of adaptation involves analyses at multiple levels of biological organization where pathways of causal determination are imbedded in networks of gene interaction and ecological interaction. It is clear that these networks are very difficult to penetrate and to unravel in a causal sense.

## SUMMARY

Initial conditions can play a determining role in the evolution of systems of knowledge just as they play a detemining role in other systems (e.g. Ruse this volume). In the case of population genetics the initial conditions focussed on the analysis of the transmission of abstract entities (genes) subject to highly simplified dynamical systems. This strategy provided the mathematical argument that was required for the reconciliation of Darwinian selection and Mendelian heredity, but it also guided and determined subsequent empirical research. This rich theoretical structure has been essential to the development of a rigorus hypothesis testing framework and it has been especially valuable in areas where historical inference is the paramount objective.

Theory has been less successful when we move from the abstract to the concrete. Thus for example, an empirical understanding of selection is much richer than implied in the simple summary parameters of population genetic theory. This is especially apparent in the study of adaptation where the essential problem is to unravel a complex system and by so doing to understand its internal workings. In contrast, the strategy of model building is to simplify by ignoring internal complexity for the sake of generality. Both approaches are valid and important, but both also circumscribe knowledge and impose real limits. One thesis of this article is that important limits to knowledge lie in the region of disjuction between theory and empirical research in population genetics.

I argued above that the study of adaptation involves a translation between different levels of biological organization - the ecological, the developmental and the genetic levels. Moreover, the study of adaptation involves spatial interactions (e.g. pollinator flight movements in the transmission flower color genes) and long temporal pathways that can not be observed in human time scales (e.g. the elaboration of gene family diversity and divergence with respect to chalcone synthase genes). Given this complexity, it seems appropriate to ask whether it is possible to unravel the precise causal pathways through which adaptive change occurs. The reality of biological systems is not linear, instead evolutionary change is determined by networks of interaction where context may be everything. These networks may be loosely connected at the boundaries that characterize different levels of biological organization. The translation across these boundaries is very poorly understood and seldom studied in contemporary biological science. We do not know how knowledge may be limited by the multiple layers of organization that exist in and are confined by a spatial and temporal context. It does seem evident, however, that the traditional methods of description are inadequate to this larger problem. Certainly future biological research will be increasingly concentrated on these boundary issues. It may be that the paradigms governing biological research will be need to be restructured in response to this imperative (as suggested by Watt and Hey in separate contributions to this volume). It is clearly evident that there are limits to knowledge imposed by the structure of current research paradigms. An exploration of these limits is an essential prerequisite to the further elaboration of evolutionary science.

## REFERENCES


Brown, A. H. D., Zohary, D. and Nevo, E. 1978. Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. Heredity 41:49-62.

Charlesworth, B., Morgan, M. T., Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular evolution. Genetics 134: 1289-1303.

Clegg, M. T. and B. K. Epperson. 1988. Natural selection on flower color polymorphisms in morning glory populations, p. 255-273. In Gottlieb, L. and S. K. Jain (eds.), Plant Evolutionary Biology. Chapman-Hall Ltd., London.

Clegg, M. T. 1997. Plant genetic diversity and the struggle to measure selection. J. Hered. 88: 1-7.

Clegg, M. T. and B. K. Epperson. 1988. Natural selection on flower color polymorphisms in morning glory populations. In, Gottlieb, L. and S. K. Jain (eds.). Plant Evolutionary Biology. pp 255-273. Chapman Hall Ltd. London.

Clegg, M. T., A. L. Kahler and R. W. Allard. 1978. The estimation of life cycle components of selection in an experimental population of barley. Genetics

Crow, J. F. and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

Cummings, M. P. and M. T. Clegg. 1998. Nucleotide sequence diversity at the alcohol dehydrogenase I locus in wild barley (Hordeum vulgare ssp. spontaneum): An evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA. 95: 5637-5642.

Dobzhansky, Th. 1981. Dobzhansky's genetics of natural populations I-XLIII. Lewontin, R.C., Moore, J.A., Provine, W. B. and Wallace, B. eds. Columbia University Press, New York.

Durbin, M. L., G. J. Learn, G. A. Huttley and M. T. Clegg. 1995. Evolution of the chalcone synthase gene family in the genus Ipomoea. Proc. Natl. Acad. Sci. U S A 92: 3338-42.

Durbin, M. L., B. McCaig and M. T. Clegg. 1999. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Molec. Biol.

Ennos, R. A. and M. T. Clegg. 1983. Flower color variation in morning glory, Ipomoea purpurea Roth. (Convolvulaceae). J. Hered. 74: 247-250.

Epperson, B. K. and M. T. Clegg. 1988. Genetics of flower color polymorphism in the common morning glory, Ipomoea purpurea. J. Hered. 79: 64-68.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection.

Fisher, R. A. 1941. Average excess and average effect of a gene substitution. Ann. Eugen. 11: 53-63.

Fu, Y.-X. and W.-H. Li. 1993. Statistical tests of neutrality of mutations. GENETICS 133: 693-709.

Glover, D., M. L. Durbin, G. Huttley and M. T. Clegg. 1996. Genetic Diversity in the common morning glory. Plant Species Biol. 11: 41-50.

Holmes, E.C., S. Nee, A. Rambaut, G. P. Garnett and P. H. Harvey. 1995. Revealing the history of infectious disease epidemics using phylogenetic trees. Phil. Trans. R. Soc. Lond. B 349: 33-40.

Holsinger, K. E. 1996. Pollination biology and the evolution of mating systems in flowering plants. Evolutionary Biology 29: 107-149.

Hudson, R 1990. Gene genealogies and the coalescent process. Oxford Surveys Evol. Biol. 7: 1-44.

Huttley, G. A., M. L. Durbin, D. E Glover and M. T. Clegg. 1997 Nucleotide polymorphism in the chalcone synthase-A locus and evolution of the chalcone synthase multigene family of common morning glory (Ipomoea purpurea). Molec. Ecol. 6: 549-558.

.

Fukada-Tanaka, S., A. HOshino, Y. Hisatomi, Y. Habu , M. Hasebe and S. Iida. 1997. Identification od new chalcone synthase genes for flower pigmentationin the Japanese and common morning glories. Plant Cell Physiol. 38: 754-758.

Kuhner, M. K., J. Yamato and J. Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140: 1421-1430.

Lewontin, R. C. 1974. The Genetic Basis of Evolutionary Change. Columbia University Press , New York.

Nee, S., E. C. Holmest, A. Rambau and P. H. Harvey. 1995. Inferring population history from molecular phylogenies. Phil. Trans. R. Soc. Lond. B 349: 25-31.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96-98.

Ohta, T. 1992. The nearly neutral theory of molecuar evolution. Ann. Rev. Ecol. Syst.

23:263-286.

Provine, W. B. 1971. The Origins of Theoretical Population Genetics.. Chicago, University of Chicago Press.

Sawyer, S. A., and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.

Simonsen, K. L., G. A. Churchil and C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Tropf, S., T. Lanz, S. A. Rensing, J. Schroder and G. Schroder. 1994. Evidence that stilbene synthases have developed from chalcone synthases several times in the course of evolution. J. Molec. Evol. 38: 610-618.

Watterson, G. A. 1975. On the number of segregating nucleotide sites in genetical models without recombination. Theor. Popul. Biol. 7: 256-276.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97-159.

Veuille, M. and L. M. King. 1995. Molecular basis of polymorphism at the esterase-5B locus in Drosophila pseudoobscura. Genetics 141:255-262.

Figure 1. A diagram of the flavonoid biosynthetic pathway.

Figure 2. A neighbor joining tree of chalcone synthase gene relationships.