Entropy **2010**, *12*, 1765-1798; doi:10.3390/e12071765

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Review

Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography

William B. Sherwin

Evolution and Ecology Research Centre, School of Biological Earth and Environmental Science, University of New South Wales, Sydney, NSW 2052, Australia; E-Mail: W.Sherwin@unsw.edu.au; Tel.:61-2-9385-2119; Fax: 61-2-9385-1558;
Homepage: http://www.bees.unsw.edu.au/school/staff/sherwin/sherwinwilliam.html

Received: 1 April 2010; in revised form: 20 June 2010 / Accepted: 28 June 2010 / Published: 15 July 2010

Abstract: This article highlights advantages of entropy-based genetic diversity measures, at levels from gene expression to landscapes. Shannon's entropy-based diversity is the standard for ecological communities. The exponentials of Shannon's and the related "mutual information" excel in their ability to express diversity intuitively, and provide a generalised method of considering microscopic behaviour to make macroscopic predictions, under given conditions. The hierarchical nature of entropy and information allows integrated modeling of diversity along one DNA sequence, and between different sequences within and among populations, species, *etc*. The aim is to identify the formal connections between genetic diversity and the flow of information to and from the environment.

Keywords: entropy, information, genes; DNA sequence; subdivision; dispersal; migration; natural selection; genome-wide association studies; linkage disequilibrium; gene expression; gene regulation; disease phenotypes

---

1. Introduction

Because of the interdisciplinary nature of this article, the appendix contains definitions of some terms and symbols. There has also been an attempt to limit jargon to terms which have workable, generally understandable definitions on web sources such as Wikipedia.

1.1. Mathematical Approaches to Predictive Biological Science

Biology is moving towards the idea that we should be able to use mathematical methods not only to analyse data from experiments or surveys, but also to make detailed predictions for outcomes of underlying processes in natural systems, or experiments. This will bring it into line with the other sciences, where mathematical modeling crept into practice during the 19th century (despite some prominent opposition), and is now entrenched. The predictive, or "front-end" mathematics starts from the behaviour of microscopic components of a system, such as molecules in a gas, allelic variants in a population of genes, or members of various species in an ecological community. Equations are then constructed which allow predictions of macroscopic properties such as pressure in a gas, range of detectable phenotypes in a population, or composition of an ecological community. These predictive equations are usually specific to certain conditions of past history (e.g., population size) and current boundaries, and require simplifying assumptions. The equations are essentially detailed hypothetical models that can be validated by simulations, and preferably by application to real-world data in which the past and present microscopic conditions are known from independent data [1]. If these hypotheses are retained, the equations may provide very explicit insights into the underlying mechanisms–the goal of science.

Despite its rather unflattering name, "back-end" mathematics is also crucial to science. It is the set of statistical methods used to assess the fit of models or hypotheses, whether they derive from front-end mathematics or from elsewhere. Where else might such hypotheses come from? For a long time, the most usual source in biology was past experience–a prediction that measure X will be higher under conditions A than under conditions B, simply because this is what had happened in the past, or through verbal combinations of past experiences with components of the system. Increasingly, this approach is being augmented by retro-fitting methods such as GLMM (Generalised Linear Mixed Models) which trial various mathematical combinations of potential driver variables, to assess their predictive power for one or more dependent variables. Note that although both past-experience and retro-fitting approaches provide numerical assessments of dependencies among variables, they offer only a broad-brush view of the underlying processes. Thus, while both front-end and back-end approaches are susceptible to the criticism that a good fit does not necessarily show causality [2,3], the lower level of mechanistic detail in the back-end approach restricts testing of causal links, as well as limiting extensions to other conditions.

This article will focus on front-end mathematics from statistical mechanics, in particular the application of entropy- and information-related concepts to genetic diversity.

1.2. Entropy, Information, and Related Approaches in Genetics

Entropy- or information-based measures of diversity provide a general tool for using microscopic behaviour to make macroscopic predictions, under given conditions. In general, the likely stable outcome is the outcome that can be achieved in the greatest number of different ways [4,5,6]. Maximum Relative Entropy has been shown to be a useful predictor for ecological community composition, giving correspondence with pre-existing fragmented theory under the same conditions [5,7], and an entropy approach has been used to assess energy flow through food-webs [8].

A huge range of methods, only loosely connected to statistical mechanics, are termed "entropy" or "entropy-like". Mathai and Haubold [9] discuss a few of these, and other genetic measures are scattered throughout their article. In some publications, the connection with entropy seems to be simply that an equation tends to be logarithmically scaled under certain conditions. Such equations include "front-end" approaches, but there are also numerous *ad hoc* "back-end" additions, with only limited verbal justification. Although I will mention the more descriptive and *ad hoc* approaches, I will focus on a small number of statistical-mechanical concepts that have been chosen for their past or increasing use in genetics, and for their potential use in predicting macroscopic behaviour from microscopic properties.

Kimura [10] and Ewens [11,12] pioneered the genetical use of the Fokker-Plank (or reverse Kolmogorov) equation from statistical mechanics. One can use this approach to model the microscopic behaviour of allelic variants in a single population at equilibrium, predicting two aspects of genetic diversity: the number of different allelic variants (richness); and the heterozygosity expected under a simple model of mutation and stochastic transmission (called "drift" by geneticists). Heterozygosity is measured as the chance that two randomly chosen alleles have a different DNA sequence:

$$H_e = 1 - \sum_{i=1}^{S} p_i^2 \tag{1}$$

where p_i is the proportion of entities of type *i* in some group (e.g., proportion of the total population of alleles that belong to type *i*, or proportions of different species in an ecological community), while *S* is the total number of extant types of alleles or species. Heterozygosity is also called by other names in genetics (e.g., haplotype diversity, nucleotide diversity) and ecology (Simpson or Gini-Simpson index) [1,7]. For a two allele system, Heterozygosity is also the binomial variance. A form of Simpson's index has been used as a null model for ecological community diversity, capitalising on the predictive power of analogous modelling in genetics [13]. In linguistics, there is a similar measure called "coefficient of coincidence" which takes various forms including $S(1 - H_e)$ [14,15].

Heterozygosity can be translated into an estimate of the number of equi-frequent elements that would be needed to give the same heterozygosity as the actual sample; this is called the "effective number of alleles–heterozygosity" [12,16]:

$$n_{eH} = 1/(1 - H_e) \tag{2}$$

Although heterozygosity and Simpson's index may not immediately seem related to entropy, below I discuss a form of Simpson's index called "Quadratic Entropy" [17].

An alternative approach to summarizing and forecasting genetic diversity can be based Shannon's diversity or entropy [18-21]:

$${}^{S}H = -\sum_{i=1}^{S} p_i \log p_i \tag{3}$$

The usual symbol *H* is modified here to avoid confusion with heterozygosity, H_e. Shannon entropy was originally proposed in the context of transmission of electronic messages, where p_i is the proportion of each of the different letters or syllables in the message. Its exponential is the number of

equi-frequent letters (or alleles, or species) that would be required to provide that same ${}^{S}H$ value. This is the "effective number of alleles–entropy":

$$n_{eS} = e^{{}^{S}H} \tag{4}$$

This is the entropic analogue of the "effective number of alleles–heterozygosity" (Equation 2). ${}^{S}H$ is the most-frequently used diversity measure in ecology [22,23]. It has seen only rare use in genetics, which is now increasing [24-29]. For both genetic and ecological uses, its shortcoming has been a lack of equations predicting its expected value under given conditions, so that until recently, it could only be a "back-end" summary. However, Ewens [11] made the first attempt to provide predictive equations, and these were extended by Sherwin *et al.* [1] to provide predictions of ${}^{S}H$ for the same conditions as Heterozygosity–a single population with random drift, and mutation. Two different types of mutation were analysed–stepwise and infinite alleles. These predictions have been validated by simulations and also showed good fit to real genetic data [1]. The predictions have not yet been extended to use with ecological community data, in the way that Hubbell [13] has transferred predictions for a close analogue of Simpson's index from genetic theory to community theory.

Each of these methods, and others which are not related to entropy, should be appraised for three features:

(a) the ability to express and partition diversity in a way that makes intuitive sense; for example when pooling K equally diverse groups which share no alleles, one might expect that the pooled diversity should be K times the diversity of each group [1,30-34]
(b) the ability to model the full range of interactions between genes, and between genes and environment in a subdivided population [1,35,36], and
(c) incorporability into a model-fitting or statistical testing framework.

2. Ability to Express Diversity in a Way That Makes Intuitive Sense

2.1. Diversity Measures and Partitioning Diversity in Genetics and Ecology

Genetics, and biology as a whole, is about diversity, but what exactly do we mean when we say that? The answer depends to great extent on the reason for which we are quantifying the diversity. Despite great interest in this area, there are surprisingly few examples of connection between diversity and function. Polley *et al.* [37] and Boyero *et al.* [38] discuss the question of whether diversity actually makes any difference to the function of ecological communities. In conservation genetics, Westermeier *et al.* [39] showed that manipulating the genetic diversity of a failing population of prairie chickens increased the recruitment rate. Moreover, diversity of sequences in humans and their pathogens has important implications for medicine, as we discuss below. Each application of diversity measures might have different mathematical requirements, depending on its intent, but wherever possible, we should seek a unified approach. Jost [30,31] pointed out that we need a diversity estimate that: (a) increases in an intuitively obvious way when one adds two different areas with the same diversity but no shared species and (b) responds in an expected way to changes, for example, more mutation or speciation yields more diversity.

Stirling [40] identified three basic components of diversity; irrespective of the relative hierarchy of these measures, each carries important information about diversity, so I will examine each component in this review:

Variety–"the number of categories into which system elements can be apportioned" [40]. This is also called "richness" in biology, e.g., the number of different allelic types or the number of different species, termed *S* in this article.

Balance–"a function of the pattern of apportionment of elements across categories" [40]. This is based on p_i–e.g., the proportions of each different type of allele.

Disparity–"the manner and degree in which the elements may be distinguished" [40]. This has been given a large number of names in biology, some of which will be introduced later.

There are measures that combine variety and balance, with different emphasis. There is a range of mathematically related measures with higher or lower emphasis on rare and common entities [1,31,41-44]. Equations (1) and (3) show that Shannon's and Simpson's (Heterozygosity) indices both include variety (*S*) and balance (p_i). Shannon's index is more sensitive to variety (*S*) than is Heterozygosity [1,30,43,44]. In each case, it is often best to convert to a measure which is "the number of equally frequent entities that gives the measured value of the diversity index" (Equations 2 or 4). Jost [30] states that the overall best diversity measure, without undue emphasis on rare or common entities, is the exponential of Shannon's index, or the number of equally-frequent entities that would be needed to provide the same diversity of information.

"Evenness" is a transform of one of these indices to make explicit the departure from the most diverse case: equal numbers of each type of allele, in which case we expect ${}^{S}H = \log S$. Evenness is:

$$E = {}^{S}H \Big/ \log S \tag{5}$$

or in the exponential scale [45,46]:

$$E' = e^{{}^{S}H} \Big/ S \tag{6}$$

Evenness can also be expressed with Heterozygosity and its derived allelic richness (Equations 1 and 2) [45]. Jost 2010 [45] showed that richness is composed of two independently varying components: evenness (*E'*) and balance expressed as n_{eS} (or their equivalents for heterozygosity). Thus two of Stirling's [40] components of diversity–richness (variety) and balance–cannot be considered to be independent, because balance is a component of richness. There has not been further consideration of the relationship of these to Stirling's other component, disparity.

For some purposes, it is reasonable to focus on diversity at one chosen level. But when we also need to account for diversity at another level, disparity becomes important. For example, when assessing diversity within a population, one might wish to somehow weight this by how different the alleles are–their "disparity" to use Stirling's term above. The disparity between the different allele types depends upon the diversity in the bases that make up each allele's DNA sequence. Thus, a population with 100 copies of each of two alleles that differ by one base, and another population with 100 copies of each of two alleles that differ at 13 base positions would both have allelic ${}^{S}H = 1$ (or $H_e = 0.5$) but we might consider that the second population was more diverse, because of the diversity at the base-level which

creates greater disparity between the two alleles. In general, there is a series of levels of diversity to consider, including levels of:

DNA sequence diversity and linkages along DNA in the genome;
Sequence diversity between different alleles within one individual, for organisms with more than one genome (e.g., diploids such as humans);
Diversity of alleles within one population;
Diversity of allele proportions in different populations of the same species;
Diversity of interactions between genes and environmental factors;
Diversity of genetics, morphology, *etc.*, between species;
Diversity of types of species within a community;

and so on.

When combining diversity measures over many levels, the tendency is to make some *ad hoc* additions to the diversity measure itself, often using qualitatively different mathematical tools at different levels. Of course, what would be most useful is to develop ways of making diversity summaries which work seamlessly across all these levels, using similar equations. Progress towards that goal will be one major theme of this review.

2.2. "One-part" Diversity Measures

It is relatively simple to integrate across multiple levels that are basically comparable. An example is to consider the allelic diversity within a population, and diversity between two sub-populations of the same species. For heterozygosity, Wright [47] defined F_{ST}, which has since been calculated in many ways, but can be seen as a function of the average heterozygosity in the two subpopulations, H_S and the heterozygosity when considering the two as a single population, H_T:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{7}$$

However, this index has a number of shortcomings, in particular, the diversity between populations F_{ST} is heavily dependent upon the allelic richness *S* within subpopulations [1]. Jost [32] points out that F_{ST} is only very weakly related to the differentiation between populations, and produced a related index which does not suffer from most of the identified shortcomings of F_{ST}:

$$JostD = \frac{H_T - H_S}{1 - H_T}\frac{n}{n-1} \tag{8}$$

where *n* is the number of subpopulations being considered. In a meta-analysis, F_{ST} and *JostD* measures were often similar [48].

Shannon's index has the advantage of being completely hierarchical, so that each level of diversity can be nested within the next, and the value of one is not affected by the other. Table 1 shows this for the case of two subpopulations, with allelic variants at a single locus in the genome, *X*. ${}^{S}H$ can be calculated following Equation 3 for individual subpopulations (${}^{S}H_1$ & ${}^{S}H_2$) or for the summed population (${}^{S}H_U$). The differentiation between the subpopulations can then be expressed by subtraction to give what is called the allelic Mutual Information:

$$MI = {}^{S}H_{U} - r_1 {}^{S}H_1 - r_2 {}^{S}H_2 \quad (9)$$

where $r_1 = N_1/N_{tot}$ and $r_1 = 1\text{-}r_2$ (*MI* was called ${}^{S}H_{UA}$ Sherwin *et al.* [1], also see Equation 9a by Lande [7]). Mutual information is more formally defined in the Glossary (Appendix 1) and in introductory texts [19-21], but it can be appreciated as the information that one variable supplies about another: mutual information between allelic identity and population membership is high when knowledge of one provides good information about the other. For example, in the two-population case, if all individuals in population 1 have allele *Xa*, and all in population 2 have allele *Xb*, then *MI* is maximal at $MI = {}^{S}H_{U}$. In other words, all diversity is between subpopulations, and there is no diversity within either subpopulation: ${}^{S}H_1 = {}^{S}H_2 = 0$.

Table 1. Entropy and Mutual information of allelic diversity in two subpopulations. Each cell entry is a frequency *f* calculated from *N* the total number of individuals in the subpopulation at location 1 or 2, and *p* the proportion of the particular allele in that subpopulation.

Location	**Allele**		**Marginal Total**
	Allele *Xa*	**Allele *Xb***	
1	$f_{a1} = N_1 p_{a1}$	$f_{b1} = N_1 p_{b1}$	$N_1 = r_1 N_{tot}$
2	$f_{a2} = N_2 p_{a2}$	$f_{b2} = N_2 p_{b2}$	$N_2 = r_2 N_{tot}$
Summed Populations, or Marginal Total	$N_a = N_1 p_{a1} + N_2 p_{a2}$	$N_b = N_1 p_{b1} + N_2 p_{b2}$	$N_{tot} = N_1 + N_2$

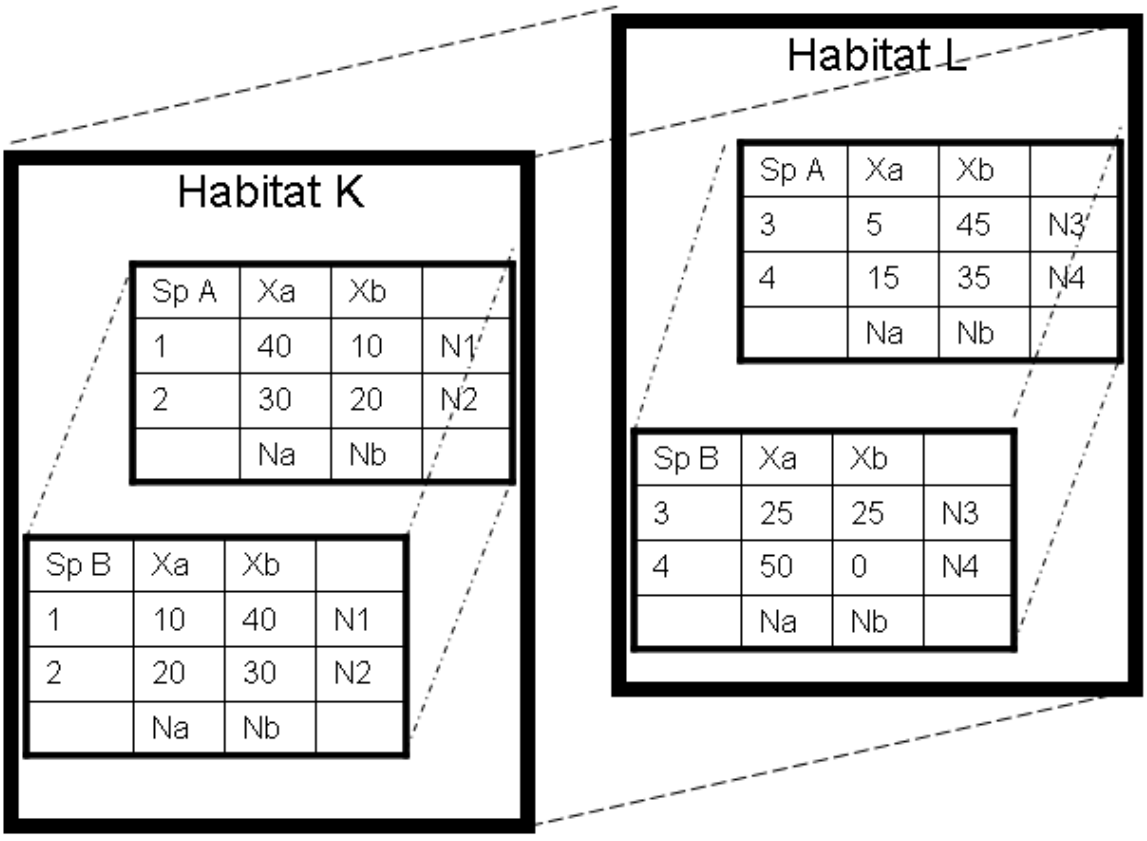
Lande [7] suggested that partitioning of diversity should utilize a contingency approach. Calculation of mutual information is made easy when it is realized that mutual information based on the *p*-values in Table 1 is linearly related to the log-linear chisquare, *G*, for a contingency test of the f-values in the same table: $G = 1.3863\ MI \times N_{tot}$ when working in Shannon's original log 2 scale [1].

If considering only two levels, there is alpha diversity within one location, and various definitions of beta and gamma diversity between locations [49]; however, there is no reason to be restricted to only two levels with entropy-related statistics. The contingency-table structure (Table 1) makes it clear how to incorporate diversity at one level as disparity at the next hierachical level of biological organization. We can add any number of dimensions to Table 1–as many as there are levels of diversity to consider–so that the table can be extended upwards from the single population more or less indefinitely, to incorporate diversity within and among different habitats, landscapes, *etc.*, each information layer becoming the measure of disparity at the next hierarchical level. Because *G* and *MI* are completely additive, *MI* between populations (Equation 9) is unaffected by allelic richness within populations, a property it shares with *Jost D* (Equation 8) [32], but not with F_{ST} [1]. Thus Shannon entropy and mutual information offer a possible route for integrating diversity across all levels of biological organization [25]. Here we are only concerned with using the contingency framework to estimate the size of various components of diversity. The alternative approach of performing a statistical test is discussed later, in Section 4.

Note that the contingency structure automatically accounts for different relative sizes of the populations at the two locations (r_1 and r_2). This property might be very important for some applications. For example, the overall biodiversity in a system that has one population with 1 *Xa* and 200 *Xb*, and another with 1 *Xa* and zero *Xb* alleles, might be considered to be quite different to the overall biodiversity of a similar system in which the second population still contains only *Xa* alleles, but has 200 of them. However, if the relative sizes of the populations are not known (as is frequently the case), r_1 and r_2 can be set to be equal. This is in fact a common assumption of most of the methods described in this paper, entropic or otherwise.

In a structure such as Figure 1, MI can be calculated for any desired interactions, using standard log-linear contingency-tests, as available in most common statistical packages [50,51]. The multiway contingency table can be interrogated by models investigating two-way interactions between any of the variables. By choice of appropriate analytical model, one can also calculate MI for interactions between higher groupings of variables. For example, a three -way interaction between alleles, habitats and species, would be identified because, in habitat K, the proportion of allele *Xa* is higher in species A (*Na* = 70, out of 100 alleles) than B (*Na* = 30, out of 100 alleles), but in habitat L, *Xa* is higher in species B than in A.

Figure 1. Generalisation of Table 1 to incorporate allelic variation (*Xa*, *Xb*) in two different locations (1,2 or 3,4) in each of two species (A, B) in two different habitat types (K,L).



When constructing models of the partition of diversity, there are always some limitations. Some variables are stratified; for example, locations 1 and 2 do not occur in habitat L, where instead locations 3 and 4 occur, therefore one would not include a location-by-habitat term in the model, or in higher interactions that included those two variables. Thus the G-value for interaction of alleles, habitat and species above would be based on the sums over locations for each species (*Na Nb*) in each quadrant of Figure 1. Also, the G-value for the interaction between species, locality, and allele would use the full data for either habitat K or habitat L, but could not combine the data for the two habitats.

There will also be cases where attributes are missing, either because they were not measured, or for some good biological reason [52]. To illustrate the latter, one could extend the dimensions of Figure 1 by duplicating it to represent allelic variation at a second locus *Y*, with alleles *Ya*, *Yb* and *Yc*. There are standard ways of summing the information from the locus-*Y* table with the information from the table for locus *X* [53]. However, it is possible that species B in habitat L simply does not have the *Y* locus in its genome–it has lost the chromosomal segment that codes for the alleles of locus *Y*. The same problem arises when considering morphological information–for example in analyzing lizard diversity, there will be no toe-length measurements available for a legless-lizard species. Of course, the presence and absence of the locus or trait is itself informative, and wherever possible this information should be represented as rows in the table whose cells contain zeros. Where the aim is to simply calculate the contribution of each interaction, as proposed here, these empty cells provide important information about differentiation. If significance testing is desired in such a situation, one would have to use methods that allow for "structural zeros"–an approach that is appropriate when there is a real-world reason why a part of a table would never ever have any entries [54].

As was pointed out in 1.1 and 2.1 above, it is often best to use an exponential transform to convert ${}^{S}H$ to a measure which is "the number of equally frequent entities that gives the measured value of the diversity index". This can be extended to the exponential of the mutual information; because of the additivity embodied in Equation 9, the exponentials will be multiplicative [30,31]. The exponential of the mutual information (and closely related statistics) is one of the few measures of diversity between populations which behaves in an intuitive manner under most situations, for example always increasing with addition of new unshared alleles, irrespective of allele proportions [34,55].

2.3. Two-part Approaches to Diversity

Although missing data and structural zeros can both be surmounted, given enough time and resources, these and other considerations have led a number of authors to take a "two-part" approach to characterizing diversity, with different approaches above and below some level of subdivision. If this is done, then detailed modeling can be complex because of the break between the mathematical treatments above and below the specified level. Therefore, these methods have not yet been subjected to modeling from microscopic to macroscopic levels, with the consequence that only verbal predictions can be made. Nevertheless, science must go on, while the theory tries to catch up, so these methods are useful.

The first such measure was a generalisation of Simpson's index/Heterozygosity, called quadratic entropy [17]:

$$Q = \sum_{i=1}^{S} \sum_{j=1}^{S} \delta_{ij} p_i p_j \tag{10}$$

where δ_{ij} represents some estimate of difference between the types (e.g., difference of morphology of species, or number of non-shared bases between alleles). The term δ_{ij} is an example of what Stirling [40] called "disparity". When all δ_{ij} values are 1, Q is equal to Simpson's index. The important thing to note is that the measurements used to calculate δ_{ij} do not have to be qualitatively similar to the measurements used to calculate p_i and p_j , freeing the investigator from many restrictions, including missing data and structural zeros. An example of a δ_{ij} distance might be one based on the number of mismatched bases between pairs of alleles (sometimes called the Hamming index, [56]) which is not easily characterisable on the same scale as p_i. Q and a related measure can be decomposed into an ANOVA (analysis of variance) approach, giving something analogous to the treatment in Figure 1 [57]. For molecular data, the program AMOVA (analysis of molecular variance) provides such an approach [58]. On the downside, the summarization of differentiation into a single distance measure misses much important information, which is why in phylogenetics, distance-based approaches are usually used not at all, or in parallel with computationally-intensive methods that use more of the information in the dataset [59]. Hardy and Jost [60] pointed out a number of other problems of Rao's Q. Ricotta and Szeidl [61] show that Q can be converted to an estimate of the number of equally abundant and maximally dissimilar species that would give the same Q as the actual value $\hat{Q} = 1/(1-Q)$.

Allen *et al*. [62] have devised a Shannon-based equivalent to Q:

$$H_P = -\sum_{b} \ell(b)\, p(b) \ln p(b) \tag{11}$$

where summation is over all branches b of a phylogenetic tree, and disparity is represented by branch lengths $\ell(b)$ that are measured on some scale such as number of substitutions in the DNA sequence, and $p(b)$ is the proportion of individuals in the present-day community who are descendants of branch b. Q and H_P will often be similar [62]. Guiaşu [63] had previously proposed this "weighted entropy" for other purposes, and analysed its properties under various conditions. In particular, the most useful form seems to occur when the weight of a branch leading to two taxa A and B $\ell(AB)$can be calculated as follows:

$$\ell(AB) = \frac{\ell(A)\, p(A) + \ell(B)\, p(B)}{p(A) + p(B)} \tag{12}$$

This may not be true of branch lengths for some types of phylogenetic analyses. Note however that Guiaşu [63] showed that this property was not necessary for all applications, including calculations of maximum weighted entropy.

Pavoine *et al*. [64] used a form of the generalized diversity equation [41-44] to show that phylogenetic weighting could be applied to any of the major indices: richness, Shannon, and Gini-Simpson. They used a variety of other weightings as well, for such things as sample size. In the light of Guiaşu's [63] findings about weights, Pavoine *et al.*'s approach probably needs further rigorous statistical analysis.

Also, Ricotta and Szeidl [61] show links between Q and H_d - a measure similar to H_P:

$$H_d = -\sum_{i=1}^{S} p_j \ln\left(1 - \sum_{i \neq j}^{S} \delta_{ij} p_i\right) \quad (13)$$

However, the different indices were not exactly equivalent. In trials on a bat dataset, with data on abundance and on divergence at mitochondrial and nuclear genes, H_P was maximized by a conservation approach that retained all species, whereas maximization of Q and H_d was achieved by elimination of rare species [62].

There are other similar measures, often with conservation applications [65] including one that is based only on variety or richness, which would be appropriate if relative abundances p_i are not relevant to the question at hand [66].

Cadotte *et al*. [67] extend H_P to incorporate abundance and divergence across the entire clade, not just for pairwise comparisons. They create an array of different diversity measures, based on verbal arguments. Several of these were called "entropy-like" because their equations are of the form $x\ln x$. Of these, their *H-AED* is the closest equivalent to H_P. In one plant data set, the entropic-like measures actually give similar results to a simple Shannon approach, but it is possible that the measures will give contrasting information to one another, when applied to other datasets. Such a result would open the way for use for these different indices to reveal different aspects of diversity. It would be useful to see microscopic modeling of the underlying genetics and ecology, to assess the properties of each of these indices.

In fact, all of these "two-part" approaches to incorporating divergence are known, or highly likely, to worsen problems for intuitive partitioning of diversity [60]. Therefore, they would benefit from formal modeling of underlying processes, and multiple or replicated tests of the work in cases where the underlying processes are known, so that the observations can be compared with theoretical expectations.

Intuitively, the theoretical modelling might well be easier for cases where the same mathematical approach is used above and below the specified level, rather than the "two-part" weighting approaches which are rather ad-hoc. In fact, the two-part methods are now facing strong competition from by joint estimation methods that use all the information simultaneously. Traditional phylogenetic reconstruction steps begin by aligning sequences with inclusion of "indels" (insertions or deletions in one or more of the sequences) to allow better alignment, followed by either using the sequences directly in character-based reconstruction methods, or via estimating disparities (distances) such as the mismatch frequency (or Hamming index) which then forms the basis of the evolutionary reconstruction [59]. Of course, this staged approach is not ideal, because the generation of indels is an important part of the real evolutionary process, and how they are added during the alignment process will affect the outcome of the later, separate, stages of phylogenetic reconstruction. There is increasing interest in joint estimation methods that perform the phylogenetic analysis without first aligning the sequences. These methods perform adequately [68,69], and some of them are based on entropy/information approaches to compressibility and distortion of signals [19-21,70-72]. Some of these can deal with complexities such as horizontal transfer of sequences between branches of the phylogeny, which makes

discordant trees in different parts of the genome [71]. The most recent entropy-based methods have not yet been extensively compared with other approaches [73].

The next two sections will examine ways in which entropy-related statistics have begun to permeate analyses, ranging from DNA base sequence level, upwards through genotype, phenotype, and interaction with environment levels, thus opening the way for fully integrated entropic treatment over all levels of biological organization.

3. Integrating Genetic Diversity Measures with Natural Processes such as Selection and Dispersal

Analysis of genetic diversity focuses on the processes which create and maintain this diversity.

Generation of diversity: This is generated by mutation plus creation of new combinations of these new variants (recombination). This diversity-generation can be thought of as equivalent to generation of new species in ecological community theory [13].

Genetic drift: This refers to stochastic sampling events in transmission between generations, especially in small populations.

Selection: Selection is seen when there is higher or lower survival and reproduction ("fitness") of some genetic types, relative to others; an obvious example is when certain allelic variants cause a disease phenotype in humans.

Subdivision: This incorporates any limitation of exchange of genes, for example non-random mating such as inbreeding within certain families, or limited genetic exchange between adjacent localities.

Attempts to identify these factors at work often start with analysis of genome-wide diversity, searching for associations between some detectable phenotype, such as disease status, and underlying genetic or environmental variation. Such analyses can provide preliminary insight into potential causal relationships which require further study [74].

Much of the existing ecological and genetic theory springs from a null model based on neutrality–all entities (species or genetic variants) are equivalent in some way, so that generation of diversity, and its stochastic transmission between generations, are the only factors to be considered in (null) predictive models. At the level of ecological communities, this means that each species is assumed to have the same rate of recruitment (e.g., Hubbell [13] for heterozygosity related measures). At the level of genes within populations of one species, this means that each genotype is assumed to have the same fitness [1,11,75]. There have been criticisms of neutral theory, but we are beginning to see extensions at the ecological level [76,77], as well as many non-neutral extensions of heterozygosity-related theory at the population genetic level [78]. Moreover, neutral theory itself has had enormous utility in genetics, when properly treated as a null model or null hypothesis. I will discuss the extension of entropy/information theory to cases which are more complex than a single random-mating population with neutral alleles at independent loci.

3.1. Subdivision

The analyses in Section 2.2 are clearly appropriate for situations where there is geographic subdivision within a species, with some genetic exchange between sub-populations. Allelic mutual information can be readily converted to an estimate of the dispersal rate per generation [1]. Compared

to all other approaches for analyzing such data, this method is robust to a wide range of dispersal and population sizes. The method has been used to assess current and historical subdivision in rainforest trees [79]. Jost's D [32] fixes most of the problems with F_{st}, but Jost's D is too sensitive to the (unknown) mutation rate to be much help for general dispersal measurement [80]. This is not necessarily a bad thing for other applications, however, because a good differentiation measure should be sensitive to mutation–the generation of new variants [33].

Another type of subdivision to be considered is not geographically-based. There is often some degree of inbreeding, ie mating with close relatives or its opposite, inbreeding avoidance. Inbreeding creates familial structure within a geographically cohesive population. Mating patterns also create a difference of information content between single alleles and diploid genotypes. For example, with high inbreeding, there would be strong correlation between the pairs of alleles received from the two parents, because the parents are relatives. This creates a high number of homozygote genotypes–individuals with two copies of the same allele at diploid loci–relative to random mating expectations. This can be dealt with by conventional heterozygosity statistics, but also by entropy- based methods. The partitioning of entropy and information between allelic and genotypic levels follows a scheme similar to that shown in Figure 1 (Sherwin *et al.* 06 supplement [1]). In parallel work, Kosman and Leonard [81] provide a variety of genetic diversity measures, based on both genotypic and allelic information, with extensions to cases where there is asexual or mixed-mode reproduction. Kosman and Leonard [81] show that the various diversity measures are not generally strongly correlated with one another, echoing findings of Sherwin *et al.* [1] for a more limited array of measures, based only on entropy. Understanding why there is or is not correlation under given conditions will require formal modelling of the population genetic processes underlying each measure. Kosman and Leonard [81] also provide transforms of Shannon entropy to give different emphasis on evenness and richness. Non-random mating will affect all loci simultaneously, increasing apparent "linkage disequilibrium" between the loci, which is discussed below.

3.2. Gene Interactions

Genes can interact through their linked inheritance and through expression networks, which include gene-gene interactions as well as gene-environment interactions. The end result will be a particular detectable phenotype, such as a particular colour of feathers, susceptibility to a disease, or a high or low reproductive rate. One of the simplest gene-gene interactions is dominance between the two alleles at the same locus. For example, many human genetic diseases are caused by recessive alleles in homozygotes–individuals with two copies of the same (defective) allele. Individuals with only one copy of the defective allele do not show the disorder, because the allele's expression is prevented by the other (dominant) allele.

Attempts to scan the genome for regions that affect particular phenotypes such as disease resistance in humans, or environmental tolerances in agricultural or wild species, are sometimes termed GWAS–"genome-wide association studies", which search for regions of the genome ("target genes") that affect the phenotype either directly, or through interactions such as affecting the regulation of other genes and physical linkage on the same DNA molecule (which can be broken by "recombination"). The number of possible interactions is very large, because the typical genome has billions of bases, each with four

alternative variants (A, C, T, G), and any part of the genome might affect phenotypes. Therefore some authors have suggested that GWAS is not the best use of our resources [82], but others have pointed out that GWAS has unearthed hitherto unsuspected candidate loci for human disorders [83] As is usually the case with bitter disputes, the truth is that both approaches are useful. Moreover, entropic methods can assist either, by viewing genes and their interactions as microstates, some of which underlie phenotypes such as human disease [84]. Compared to other approaches, such as multiple regression, contingency/entropy methods may have more power to detect interaction between genetic variants which individually have no direct effects on the phenotype (*i.e.*, no "marginal effects" in a table such as Figure 1, where "Species A" and "Species B" were replaced by "diseased" and "non-diseased" phenotypes) [85].

Kang *et al.* [84] consider the entropy of a system of loci which potentially interact in determining disease status, for example, the two single-nucleotide polymorphisms (SNPs) in Table 2. If there is no interaction between the SNP loci, then the occurrence of each combination *g1*,…*g9* should be governed entirely by the proportion of each single-locus genotype (e.g., the proportion of *CC*, *TT* individuals would be *P(CC) x P(TT*). These expected proportions can be used in Equation 3 to calculate the expected entropy ${}^{S}H_{\exp}$. This entropy can be compared with the observed entropy ${}^{S}H_{obs}$, calculated from the proportions of the two-SNP genotypes in the sample, *g1*,…*g9*. A departure from independence might indicate interaction between the loci. The sample could be taken from individuals showing the disease phenotype, or from the general population. Kang *et al*. [84] subtract entropies from one another, in a manner similar to conditional or relative entropies, or Kullback-Liebler distances [19-21]. This produces $\Delta H = H_{\exp} - H_{obs}$, which is distributed like a chi-square statistic, for reasons that will be obvious, due to its similarity to Figure 1 and associated equations. Kang *et al.* [84] also calculated a ratio of entropies $I = 1 - \left(H_{obs} / H_{\exp} \right)$. A similar statistic was proposed by Smouse and Ward [25]. Kang *et al.* then created specific models of interactions with mathematically explicit effects on the two-locus genotype proportions g1,…g9. They simulated data under these models, and used the results to compare the power of ΔH with more conventional chi-square tests for interaction. ΔH showed better type I and type II errors, and also performed well on real data for two diseases–schizophrenia and malaria. Their I statistic proved to be useful for graphical display of significant results.

Table 2. Diploid individuals, with two single-nucleotide polymorphism loci (SNPs), each of which have two allelic variants, say bases *A* and *C* at SNP locus 1, and G and T at SNP locus 2, giving nine possible two-locus genotypes, in proportions *g1*,…*g9*.

		SNP 1		
		AA	*AC*	*CC*
SNP 2	*GG*	*g1*	*g2*	*g3*
	GT	*g4*	*g5*	*g6*
	TT	*g7*	*g8*	*g9*

Similarly, Dong [86] calculated mutual information between each of two SNPs and disease status, then showed that there is an additional gain of information from considering interaction between *SNP1*

and *SNP2*. After successful use in simulations, they then tested the methods with susceptibility to malaria infection, demonstrating the expected negative epistasis between the effects of two globin polymorphisms (sickle-cell and a-thalassemia).

Chanda *et al.* [87] took on the general question of assessing potential gene-gene interactions and gene-environment interactions on a massive scale, focusing on identifying strengths of association rather than significance testing. They used a mutual-information-like approach to summarise gene-gene-environment dependencies, most notably in their *KWII* or "K-way interaction information". For three factors A, B, C there would be a three-way contingency table (Table 3). *KWII* is then calculated as:

$$\mathrm{KWII(A,B,C)} = -\,{}^{S}\mathrm{H(A)} - {}^{S}\mathrm{H(B)} - {}^{S}\mathrm{H(C)} + {}^{S}\mathrm{H(AB)} + {}^{S}\mathrm{H(AC)} + {}^{S}\mathrm{H(BC)} - {}^{S}\mathrm{H(ABC)} \quad (14)$$

${}^{S}H(A)$, ${}^{S}H(B)$, ${}^{S}H(C)$ would each be calculated from the marginal totals in Table 3 for that variable. Thus ${}^{S}H(A)$ would be calculated using Equation 2 on the proportions of the three genotypes in the whole dataset, ie the *p*-values: 116/364, 155/364 and 93/364. ${}^{S}H(AB)$ would be calculated using Equation 2 on the proportions in a table that had been summed over all categories of factor C, giving p-values 24/364; 56/364; 36/364; 49/364; 26/364; 80/364; 23/364; 20/364; 50/364. Chanda *et al.* also use a related variable, *TCI*, which assesses only the extent to which the marginal entropies ${}^{S}H(A)$, ${}^{S}H(B)$ and ${}^{S}H(C)$ are independent of the highest-level interaction, in this case ${}^{S}H(ABC)$:

$$\mathrm{TCI(A,B,C)} = {}^{S}\mathrm{H(A)} + {}^{S}\mathrm{H(B)} + {}^{S}\mathrm{H(C)} - {}^{S}\mathrm{H(ABC)} \quad (15)$$

and finally, a measure called phenotype-associated information, *PAI*, which compares the *TCI* with and without the phenotypic variable (e.g., ${}^{S}H(C)$) to evaluate the difference in dependencies with and without the phenotypic information. In the three-way case of our example,

$$\mathrm{PAI(A,B,C)} = \mathrm{TCI(A,B,C)} - \mathrm{TCI(A,B)} \quad (16)$$

Table 3. Interaction between three factors: A one diallelic gene, B one environmental variable, and C one phenotype (diseased/non-diseased).

(a)

C-diseased		B ENVIRONMENT–temperature		
		low	*mid*	*high*
A the SNP	*GG*	*12*	*43*	*18*
	GT	*45*	*13*	*2*
	TT	*20*	*10*	*0*

(b)

C-non-diseased		B ENVIRONMENT–temperature		
		low	*mid*	*high*
A the SNP	*GG*	*12*	*13*	*18*
	GT	*4*	*13*	*78*
	TT	*3*	*10*	*50*

Where there are multiple genes, multiple environmental factors, and multiple phenotypes, there is a very large number of potential interactions, each of which can be given a value of *KWII*, and there are similarly large numbers of *TCI* and *PAI* values. These values can be visually presented in a histogram, allowing the user to parsimoniously select interactions for further study, for example, those with the largest *KWII*. Chanda *et al.* [87] trialed the method on a simulated dataset based on public data on SNP genotypes and rheumatoid arthritis, to which there had been added deliberate interactions at various levels from first-order marginals to multiway interactions. They compared its performance with three methods: two types of multifactor dimensionality reduction (MDR), and a pedigree disequilibrium test. Each method was provided with its optimal dataset from the simulated data, though it should be noted that pedigree data would typically be much more difficult to obtain than the non-pedigreed case-control data required by Chanda's method and MDR. Chanda *et al.* found that *KWII* performed adequately in comparison with the other approaches, though no method found all the known interactions that had been engineered into the dataset. One great benefit of Chanda's method seems to be its easily-used visual display, but even this becomes difficult when the method is expanded to cope with 10^6 or 10^7 loci from a typical SNP genotyping array. Thus they resort to a program, AMBIENCE, which first searches for marginal interactions (one-way interaction of a single SNP locus and disease status), then the program only analyses higher-order interactions between those SNPs, using a search strategy which could potentially miss some higher interactions, especially those which do show any marginal effects.

Of course, many diseases and other phenotypes cannot be coded "present" and "absent" as in Table 3, but have a continuous distribution which might nevertheless be different in individuals with different SNP genotypes and environmental exposure. The entropy of such a continuous distribution ("differential entropy" *h*) can be calculated by simply using continuous version of Equation 3, based on the frequency distribution for each variable, e.g., (*f(a)*) for variable *A* [88]:

$$h_A = -\int \mathrm{f(a)} \ln \mathrm{f(a)}\, \mathrm{da} = \ln\left(\sigma\sqrt{2\pi\mathrm{e}}\right) \tag{17}$$

where σ is the standard deviation of variable *A*. This depends only upon the variance of *A*, as expected, and assumes that the data have been transformed to achieve normality. An analysis of a simulated dataset based on public data on SNP genotypes and rheumatoid arthritis showed that *PAI* performed similarly to a competing analysis termed the "restricted partitioning method" [88]. Again the algorithm (CHORUS) used a search which could be entrapped by locally optimal values, so further development is needed to deal with massive datasets.

Another approach to the analysis of interactions between genes might benefit from analogy with ecology. Volkov *et al*. [89] used a maximum entropy approach to investigate the interactions between abundances of the 20 most abundant tree species at Barro Colorado Island, which had previously been analysed under simpler assumptions such as all species having equal demographic rates (*b/d* or *b-d*, where *b* is birth rate and *d* is death rate) [13]. Volkov *et al.* showed that the maximum entropy approach provided results which were comparable to a more conventional transition matrix approach which allowed for variation in recruitment rate.

The treatment of single SNPs and their pairwise interactions has been criticised as missing information that is contained in whole haplotypes–blocks of DNA containing multiple SNPs, and

coding for one or more genes and regulatory regions. Various approaches are being investigated as possible ways to identify appropriate haplotype blocks for GWAS and other purposes. Even approaches that are not directly founded on entropy, such as Bayesian methods, sometimes include components that are best termed "relative entropy" (see Appendix) [90]. Cui [91] extend beyond single SNP-disease association to incorporate all SNPs in a gene simultaneously, using an entropy measure to summarize the interaction between all SNPs and the disease state. They showed that this has greater power than an approach that deals with each SNP independently.

Interactions between loci are not only because of their effects on the phenotype. For example, a block of DNA might contain two SNPs, one with alleles A or C, and the other with alleles G or T. This stretch of DNA could then have ___A___G___ or ___A___T___ or ___C___G___ or ___C___T___ in different "haplotypes". This linkage of the SNPs means that inheritance of allelic variants at one SNP locus is not statistically independent of alleles at the other, and thus the probability of observing particular haplotypic combinations of alleles at the different loci, is not equal to the product of the allele proportions at the loci. This non-independence, resulting from physical linkage on the same DNA molecule, is often called "linkage disequilibrium", and can be eliminated by "recombination"–breakage and rejoining of DNA molecules. Linkage disequilibrium can also be eliminated or intensified by stochasticity in transmission in small populations ("random genetic drift"), which alters the proportions of haplotypes without the need for recombination. Linkage disequilibrium has become of intense interest for reasons that include: its great resolution for fine-scale gene mapping [92,93], especially in cases of admixture [94]; importance in forensics; and use in identification of regions of the genome that are under selection [95-97]. Conversely, a great many population genetic analyses assume linkage equilibrium, so the possibility of disequilibrium must be evaluated before proceeding with further analysis [94].

Apparent linkage disequilibrium can actually result from many causes, only some of which are to do with actual physical linkage. One such source of apparent disequilibrium is selective advantage or disadvantage for particular combinations of alleles at different loci (called "epistasis" for fitness), irrespective of whether the loci are linked in the same DNA molecule. Thus there are investigations of multi-locus selection on unlinked loci [98].

A variety of methods have been proposed to infer and estimate apparent linkage from population data, some methods being based on summary statistics, others on Bayesian or likelihood approaches, and some on hybrids between the two [92,93,99-102]. Zhang [103] used mutual information between pairs of loci such as the two SNP loci in Table 2, as a measure of linkage disequilibrium, and showed that it is proportional to the correlation between haplotypes in the diploid genotype r^2. Hampe [104] used entropy-based methods to deal with the very important operational problem of which SNPs to type before embarking on a GWAS. Once one or more SNPs have been chosen, the choice of which is the best SNP to add depends upon which additional SNP will give the greatest increase of power for the GWAS. The increase of power is a function of the new SNP's variability, its position relative to the SNPs already chosen, its linkage disequilibrium relative to the prior SNPs, and its position relative to target genes: regions of DNA code which directly or indirectly affect the phenotype being investigated, such as a human disease. Hampe [104] provides a method to choose the best additional marker on the basis of greatest increase of mutual information with the target gene. Of course, at this preliminary

stage of the GWAS, the identity and location of target gene(s) is unknown, or only guessed. Methods must therefore be generalized over all possible target genes [104]. The method saves up to 30% of genotyping load, compared with simply typing equidistant SNPs, or pairwise maximization of linkage disequilibrium between SNPs [104]. Shannon's index has also been used as the basis of a measure of founder informativeness for optimizing quantitative trait locus mapping with any level of inbreeding or outbreeding [105].

3.3. Selection

When different genotypes code for phenotypes with different survival or reproductive rates, we say that the genotypes differ in fitness. Essentially, GWAS is searching for the genomic regions underlying one or more aspects of total fitness. Genotypes with higher fitness will obviously become relatively more numerous, and this "selection" is the basis of adaptation to environmental conditions. Because genetic variants provide the raw material for this process, it is thought that genetic diversity is crucial for continued adaptation to the changing environment [106]. There are surprisingly few direct demonstrations that increased genetic diversity raises recruitment rate in wild populations, and these can be disputed in many cases [39,106-108]. Thus there is scope for many further studies of the interaction between environmental factors and genetic diversity, and entropy-based measures can assist here.

Selection comes in three basic forms. The first is directional selection, which eliminates one genetic variant in favour of the other, such as loss of all *A* alleles from *SNP1* in Table 2. Subtypes of directional selection include negative selection which considers the loss of *A*, and positive which considers the effect on *C*. Secondly, stabilising or balancing selection refers to various types of selection which tend to maintain variants, for example preferential survival of *AC* heterozygotes, relative to individuals with the homozygous *AA* and *CC* genotypes in Table 2, which would result in persistence of both *A* and *C* alleles. Balancing selection, of various types, is suspected in the major histocompatability loci which affect immune response, mate choices, and reproductive success in humans and other vertebrates [109]. Finally, divergent or disruptive selection is when the intermediates such as *AC* are at a disadvantage. This is clearly an unstable situation, but may occur in hybrids.

Directional selection has recently been analysed by statistical-mechanical approaches. Saakian and Fontanari [110], without explicitly mentioning entropy, use an information-based approach to analyzing selection in what appears to be a haploid system. Barton and Coe [6] used an information-like entropy, first developed by Sella and Hirsch [111], to show that with selection, a genetic system tends towards the state that can be realized in the largest number of ways, reflecting other findings in uses of entropy in biology, such as maximum relative entropy [3-5]. Barton and Coe [6] provide some discussion of whether to regard the allelic variants, or the diploid genotypes, as the microscopic entities. This could be resolved by partitioning entropy/information into genotypic and allelic components [1 supplement, 81]. Barton and Coe [6] point out that their approach is yet to be validated by simulations and natural studies. Iwasa [112] modeled a variety of situations, from selection for codon usage through to evolution of continuously-variable "quantitative" traits, and showed that these can be modeled using an equation called "free fitness" whose general form is:

$$\textit{Free fitness} = \textit{selection component} + \textit{entropy term} \tag{18}$$

The first term is always based on the mean fitness of individuals in the population ($\overline{W}$), while the second is based on the entropy of the array of allele proportions (Equation 3), and its evolution incorporates stochastic processes such as mutation and random genetic drift. Using Iwasa's formulation, Barton and de Vladar [113] showed that additive directional selection on a quantitative trait will tend to maximize relative entropy. The results strictly apply to cases with relatively high mutation ($4N_e\mu > 1$, where N_e is effective population size and μ is mutation rate), but can be extended to lower mutation rates. Within this work, there is an abundance of parallels with statistical mechanics: adaptive landscapes of fitness can be seen as analogous to energy gradients, and the entropic approach has a close analogue of the additive genetic covariance matrix [113]. Adami [35] briefly mentions a parallel approach to entropy and directional selection, in the context of selection on gene regulatory networks.

Other forms of selection can be analysed by entropy-based treatments, and at a simplistic level, they are not necessarily difficult. For balancing selection where the relative survivals of *AA*, *AC* and *CC* in Table 2 are 1-s_1, 1, and 1-s_2 $(0 \le s_1, s_2 \le 1)$, then the expected equilibrium proportion of *A* would be $\mathrm{P(A)} = \mathrm{s}_2/s$ where $\mathrm{s} = \mathrm{s}_1 + \mathrm{s}_2$, [78]. Therefore, at equilibrium, Equation 3 gives the allelic entropy for *SNP1* to be:

$$\hat{H} = \log s - (s_1 \log s_1 + s_2 \log s_2)/s \tag{19}$$

where the first component is the effect on entropy of the overall strength of selection, and the second is sensitive to the effect of the fitness-differential between the two homozygotes, AA and CC. Barton and de Vladar [113] extended their entropy-based analysis of additive selection on quantitative traits to include other interactions such as dominance and epistasis. Mustonen and Lassig [114] define fitness flux as the product of selection coefficients and rate of change of allele proportions, and show that this unifies and extends many other theories of adaptation, including Iwasa's relative-entropic approach. The predictions of the fitness-flux theory are consistent with evolution of bacteria, which evolved to a state showing increased fitness and decreased fitness flux [114].

Schwanz and Proulx [115] used entropy and information methods to model a type of balancing selection due to frequency-dependence–when the fitness of a genotype depends upon its frequency. If this dependence is negative, so that rare genotypes are favoured, then this will maintain multiple genotypes. They applied it to temperature-dependent sex determination, such as occurs in turtles. Clearly there can be frequency dependence in selection on sex-ratio. If a rare genetic variant tends to make its carriers into males at prevailing temperatures, then this variant will be at an advantage if all other variants make females at those temperatures. Half the alleles in the next generation will be provided by the few carriers of the male-determining variant. The same advantage would accrue to a rare female-determining variant, so that both male- and female-determining variants will tend to be maintained in the population. The evolution is determined by interaction between different genetic variants with different threshold temperatures for male and female determination, and the variation of the temperature regime over many generations. These authors [115] used mutual information as a measure of the interaction between genotypes and environmental conditions (called the "reaction norm" in quantitative genetics) and of the evolution of this relationship. They simulated the extent to

which temperature affects sex under different levels of environmental variation, and with different levels of overlap of generations, which affects the number of temperature regimes experienced by each generation. Their model identified that mutual information increases if a "switch-like" variant exists, with a sharp transition from male-determination to female-determination at some temperature within the natural range.

4. Ability to be Incorporated into an Inclusive Statistical Framework

There are two broad approaches to statistics: to test hypotheses, or to make estimates of parameters for particular models, with appropriate confidence limits. The latter is gaining popularity, and this review has emphasized it. For either purpose, an ideal statistical framework would allow partitioning of diversity over the full scope of living systems, from spatial structure of an ecological community through to genetic structure at the scale of the single base pair in the genome [1]. To approach this goal, diversity estimators such as F_{st} and derivatives have been used in an ANOVA type framework, and developed further as spatial autocorrelation, AMOVA [58], spatial AMOVA, *etc*. For the entropy-based measures, the candidate is mutual information. For partitioning of diversity, mutual information has two great advantages over the heterozygosity-based measures: it is completely additive over hierarchical levels, and it has a close relationship to the log-linear contingency test [1,25]. Mutual information is linearly related to the log-linear chisquare, *G*, for a contingency test of the same table (e.g., Figure 1) $G = 1.3863\ MI \times N_{tot}$ when working in Shannon's original $\log_2$ scale [1]. The user can choose whether to use the resulting values of mutual information as estimates of effects on diversity of particular combinations of factors, or for conversion to p-vales for a statistical test. As noted above, when statistical testing is desired, there are some additional restrictions on the use of the data [54].

Moreover, there is additional information available if the linkage or other associations of allelic variants at different genes (loci) are considered. Each locus might respond differently to some processes, such as fitness differences that affect carriers of an allele at only one locus. However, the loci will behave similarly under other processes, such as non-random mating, or immigration, which affect all loci. Therefore there have been attempts to unite linkage (or gametic) disequlibrium with spatial genetic structure measures such as F_{st} [116-118]. However, the hierarchical structure of the entropy-based measure mutual information makes it particularly powerful in this regard [25,119]

The small number of practitioners of genetic entropy analyses have been relatively good at actually validating the methods that they propose. Three levels of validation are used. Firstly, simulations allow good replication and controls, but can never incorporate all the complexities of real living systems, which are, after all, the desired use-point. Secondly, there are sometimes implementations with data from real populations. These include cases where the observations can be compared with expected results forecast from a knowledge of the underlying driver variables [1,120]. However, in comparison to simulations, these typically have poorer replication and sometimes few or no controls. Finally, the entropy statistics can be calculated from observations in real populations, but without any independent data that can be used to create expectations for the entropy statistics. Such analyses are quite common, and are important to show that the methods can be implemented with real data, but cannot be said to be a critical test of the method in the formal scientific sense, and most authors are careful not to make such a claim.

How good does an entropy-based approach need to be to become acceptable? Chanda *et al*. [87,88] used simulations and real data to assess the power of their entropy-based method, relative to contenders, and concluded that its statistical power was "acceptable"–probably too weak a statement to get a new method widely used. However, entropy-based methods have another point in their favour. Most of the competing analyses are *post hoc* methods–such as regression [85]–which can identify association, but which lead to less insight into the underlying mechanisms. This is because, unlike entropy, they are not firmly based on a model of underlying microstates–the network of interacting genes and environmental effects. It is up to the researcher to choose the level to treat as "microscopic". For various applications, alleles, individuals, populations, or species might be appropriate microscopic levels. Of course the hierarchical nature of entropy and information systems means that these can be nested within one another almost seamlessly.

It should be noted that all measures in this article, entropic or otherwise, can suffer from estimation problems when samples are small [121,122]. The correction for number of effective alleles n_{eH} (Equation 2) is in Neilsen *et al.* [123], and for other non-entropy measures, corrections are available in introductory texts [78]. For entropy-based statistics, small sample corrections are also available [124]. Chao and Shen [125] have further developed the theory of sampling for Shannon diversity.

5. Future Directions

In some ways, the data are considerably ahead of the theory–people can collect data for which there are no theoretical predictions, except rather broad-brush verbal suggestions. In other ways, the theory is considerably ahead of the data: in many cases, theory is already available (and waiting to be tested), but the appropriate data have yet to appear. The necessary data may come soon with improvements in high-throughput genetic techniques. Again and again, entropy-based approaches emerge as a solution to a problem in analysis of genetic variation and its interactions at all scales.

For an overall integrated approach to genetic modelling via entropy, a very important field is the integration of variation along a DNA sequence, with variation between different sequences within and between populations, species, landscapes *etc*. There have been descriptive uses of entropy in DNA sequence analysis, including: genome organization [126]; synonymous codon usage biases in humans and mice [127]; viral DNA sequences [56,128]. However, what will be ultimately needed are methods that formally connect diversity along DNA sequences to the flow of information to and from the environment, thus making a predictive framework for the way that genetic diversity interacts with environmental diversity. This work therefore will have strong connections to network theory, as reviewed by Gatenby and Frieden [36].

The field of modeling and analyzing diversity along DNA sequences is very much in its infancy, but there have been some interesting applications of entropy theory. Schneider [129] used entropy calculations to analyse the information contained in a particular DNA region, such as a splicing sites which must be recognized against the background of the other information in the entire genome. Schneider [129] considers the amount of information that the site must contain to differentiate it from all others in the genome, and shows how this can evolve to a predictable value. Schneider [129] also shows correspondence between rate of evolution as bits per generation, and the (rarely used) evolutionary measures of Darwins or Haldanes, both logarithmic time-and-rate scaling conventions.

In a very clear review, Adami [35] made a case that entropy, or the information derived from comparison of equilibrium and non-equilibrium states, could be a useful tool for making predictive models of DNA sequence evolution, including prediction of protein structure from the sequence of the DNA code. However, Adami did not actually evaluate whether the method works better than other approaches. Sanchez and Grau [130] extend the idea of DNA's triplet code or amino acids as a Boolean information system, and from this suggest that the distribution of amino acids in a protein might follow a Boltzmann distribution. Long ago, Iwasa [112] wrote theory for evolution of codon usage bias in entropic terms, and this has recently been rediscovered in analysis of codon usage in yeast [131]. Loewenstern and Yianilos [132] show that the entropy of natural DNA is less than its theoretical maximum of 2 bits per site.

There are some cases where entropy and information theory have allowed us to improve our understanding of the phenotypic effects of particular DNA sequences and their mutations. In the genes for two human blood-clotting factors, there are various mutations that affect the splicing of RNA that carries the code to the cell to be converted to the protein that makes the clotting factors. Some mutations result in excessive bleeding in the patient, but it is not logistically possible to examine all the possible mutations [120]. However, information analysis of splicing sites has been used to predict accurately which mutants would lead to clinical symptoms [120].

Entropy and information theory has been used several times in analysis of complex regulatory networks of genes and environmental signals. Diaz *et al*. [133] have used entropy methods to model the behaviour of *ERF*, a master gene which regulates plants' response to ethylene, and the consequent behaviour of the ethylene signaling pathway. Adami and Hintze [134] have extended this to incorporate multiple nodes in the information pathway in an artificial evolution situation, and showed that fitness, information, and number of nodes all increase. Presumably there is some limit to this increase, due to energetic considerations or other tradeoffs. Lezon *et al*. [135] used entropy maximization to infer the presence of strongly interacting pairs of genes in regulatory networks. They claimed that the method preferentially identified master-slave gene pairs in preference to pairs of slaves with the same master, but a direct comparison with alternative ways of making such identifications has yet to be constructed.

There is increasing incorporation of entropy methods in user-friendly genetic analysis platforms, so that entropy-based measures can now be readily estimated from datasets. Shannon information and mutual information are now available in GENALEX 6.3 [136]. Given data transferability between programs, entropy methods can therefore be used with genetic data sourced from most other shareware population genetic analysis programs. Hickerson has incorporated Shannon information and mutual information into MSBAYES, an Approximate Bayesian Computation program for molecular evolution [137]. This program chooses various statistics, depending on their ability to create rapid convergence to a solution. Other programs which include entropy statistics for summarizing genetic diversity include MSA [138] and a program aimed at analysis of polyploid genotypes [139,140].

Entropy- and non-entropy-based measures might also be used together, to capture different aspects of the data. Pielou [141] suggested displaying diversity as a plot of each type of diversity index, for example, the orders of Hill's [42] generalized index (whose powers include richness, and Equations 1 and 3). However, note that this comparison can give starkly contradictory results depending upon

whether the researcher uses the additive measures such as Equations 1 and 3, or the multiplicative measures such as Equations 2 and 4 [34]. Chao [55] has generalized the Morisita-Horn index, which is also closely related to Simpson's index, and Jost *et al*. [34] have shown that this could form an important link between the different indices, as well as going some way to reconcile additive and multiplicative approaches. Additionally, the entropic version of the Horn index allows incorporation of weighting by absolute abundances [142], which may be a very useful property, as discussed above for MI.

Finally, in the same way that the predictive power of other population genetic analyses is being transferred to ecology [13], the predictive power for entropy-based approaches derived in population genetics [1] could be transferred to ecology.

6. Summary

The entropy and information biodiversity measures are the most-frequently used diversity measures in ecology, but until recently have been used only sporadically in genetics. They promise to provide explicit mechanistic insights by allowing modeling of microscopic behaviour to make macroscopic predictions. Using entropy/information theory, predictive equations have recently been developed for various genetic systems ranging from a single locus in a single population with random drift and mutation, to continuously variable traits with selection of various types.

Entropy and information methods are particularly good at partitioning data to investigate effects. Mutual information is obtained when entropy is partitioned, for example between geographic regions or between genetic or phenotypic classes. This has been used for a wide range of purposes including: making very robust estimates of the dispersal between populations, investigating mating patterns, analysing gene-gene-environment interactions in production of phenotypes such as disease states, and analysing linkage between genes. When partitioning genetic diversity, the exponentials of entropy and information are some of the few measures of diversity which behave in an intuitive manner under most conditions.

The hierarchical nature of entropy and information allows integrated modeling of diversity across all levels of biological organization: gene regulation; prediction of protein and disease phenotypes; genome organization such as codon usage; variation along one DNA sequence; variation between different sequences within and among populations, species, *etc*. Thus, we can model the full range of interactions between genes, and between genes and environment in a subdivided population. There needs to be further exploration of the flow of information to and from the environment and the genome.

Entropy approaches are easily incorporated into a model-fitting or statistical testing framework, and have been especially useful for efficiently dealing with the huge numbers of potential interactions in genome-wide association studies of human diseases. However, further work is needed to improve their reliability and efficiency. There is increasing incorporation of entropy methods into user-friendly genetic analysis platforms.

The predictive power for entropy-based approaches to genetics might well be usefully extended to ecology and other fields of science.

Acknowledgements

My work in this area has benefited greatly from many discussions with colleagues including Roddy Dewar, Franck Jabot, Mike Hickerson, Lou Jost, Michael Lachmann, Dennis McNevin, John Morris, Dhriti Pandya, Rod Peakall, Murali Ramanathan, Maurizio Rossetto, Jurgis Sapijanskas, Claire Sadler, Peter Smouse, Alan Welsh, and the members of my lab. In particular, Roddy Dewar, Lou Jost and Peter Smouse made detailed comments on an early draft of the manuscript. I also thank the anonymous reviewers and Peter Harremoës for their very useful comments.

References and Notes

1. Sherwin, W.B.; Jabot, F.; Rush R.; Rossetto M. Measurement of biological information with applications from genes to landscapes. *Molec. Ecol.* **2006**, 15, 2857-2869.
2. Zar, J.H. *Biostatistical analysis*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1984.
3. Banavar, J.R.; Maritan, A.; Volkov, I. Applications of the principle of maximum entropy: from physics to ecology. *J. Phys.: Condens. Matter* **2010**, *22*, 063101.
4. Dewar, R.C.; Porté, A. Statistical mechanics unifies different ecological patterns. *J. Theoret. Biol.* **2008**, *251*,389-403.
5. Dewar, R.C. Maximum entropy production as an inference algorithm that translates physical assumptions into macroscopic predictions: Don't shoot the messenger. *Entropy* **2009**, *11*, 931-944.
6. Barton, N.H.; Coe, J.B. On the application of statistical physics to evolutionary biology. *J. Theoret. Biol.* **2009**, *259*, 317-324.
7. Lande, R. Statistics and partitioning of species diversity and similarity among multiple communities. *Oikos* **1996**, *76*, 5-13.
8. Zhang, J. Modeling multi-species interacting ecosystem by a simple equation. *Int. Joint Conf. Comp. Sci. Opt.* **2009**, *1*, 1003-1007.
9. Mathai, A.M.; Haubold, H.J. On generalized entropy measures and pathways. *Phys. A* **2007**, *385*, 493-500.
10. Kimura, M. Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symp. Quant. Biol.* **1955**, *20*, 33-53.
11. Ewens, W.J. The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **1972**, *3*, 87-112.
12. Ewens, W.J. *Mathematical Population Genetics*; Springer-Verlag: New York, NY, USA, 1979.
13. Hubbell, S.P. *The Unified Neutral Theory of Biodiversity and Biogeography*; Princeton University Press: Princeton, NJ, USA. 2001.
14. Friedman, W.F. *The Index of Coincidence and its Applications in Cryptology*; Riverbank Laboratories, Department of Ciphers: Geneva, IL, USA, 1922.
15. Index of coincidence. http://en.wikipedia.org/wiki/Index_of_coincidence (accessed on 8 July 2010)
16. Kimura, M.; Crow, J. The number of alleles that can be maintained in a finite population. *Genetics* **1964**. *49*,725-738

17. Rao, C.R. Diversity and dissimilarity coefficients: a unified approach. *Theoret. Pop. Biol.* **1982**, *21*, 24-43.
18. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379-423, 623-656.
19. Bell, D. *Information Theory*; Pitman: London, UK, 1968.
20. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
21. Csiszár, I.; Shields, P. Information theory and statistics: A tutorial. *Found. Tr. Commun. Inform. Theor.* **2004**, *1*, 417-528.
22. Buddle, C.M.; Beguin, J.; Bolduc, E.; Mercado, A.; Sackett, T.E.; Selby, R.D; Varady-Szabo, H.; Zeran, R.M. The importance and use of taxon sampling curves for comparative biodiversity research with forest arthropod assemblages. *Can. Entomol.* **2004**, *137*, 120-127.
23. Bulit, C.; Diaz-Avalos, C.; Montagnes, D.J.S. Scaling patterns of plankton diversity: a study of ciliates in a tropical coastal lagoon. *Hydrobiologia* **2009**, *624*, 29-44.
24. Lewontin R.C. The apportionment of human diversity. *Evol. Biol.* 1972, *6*, 381-398.
25. Smouse, P.E.; Ward, R.H. A comparison of the genetic infra-structure of the Ye'cuana and Yanomama: A likelihood analysis of genotypic variation among populations. *Genetics* **1978**, *88*, 611-631.
26. Hartl, G.B.; Willing, R.; Nadlinger, K. Allozymes in mammalian population genetics and systematics: Indicative function of a marker system reconsidered. *Exp. Suppl.* **1994**, *69*, 299-310.
27. Lacerda, D.R.; Acedo, M.D.P.; Lemos Filho, J.P.; Lovato, M.B. Genetic diversity and structure of natural populations of *Plathymenia reticulata* (Mimosoideae), a Tropical Tree from the Brazilian Cerrado. *Molec.Ecol.* **2001**, *10*, 1143-1152.
28. Wang, T.; Su, Y.J.; Li, X.Y. Genetic Structure and Variation in the Relict Populations of *Alsophila spinulosa* from Southern China based on RAPD Markers and cpDNA atpB-rbcL Sequence Data. *Hereditas* **2004**, *140*, 8-17.
29. He, T.; Krauss, S.L.; Lamont, B.B.; Miller, B.P.; Enright, N.J. Long-distance seed dispersal in a metapopulation of *Banksia hookeriana* inferred from a population allocation analysis of amplified fragment length polymorphism data. *Mol. Ecol.* **2004**, *13*, 1099–1109.
30. Jost, L. Entropy and diversity. *Oikos* **2006**, *113*:363-375.
31. Jost, L. Partitioning diversity into independent alpha and beta components. *Ecology* **2007**, *88*, 2427-2439.
32. Jost, L. Gst and its relatives do not measure differentiation. *Mol. Ecol.* **2008**, *17*, 4015-4026.
33. Jost, L. D *vs.* G_{ST}: Response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Mol. Ecol.* **2009**, *18*, 2088-2091.
34. Jost, L.; DeVries, P.; Walla, T.; Greeney, H.; Chao, A.; Ricotta, C. Partitioning diversity for conservation analyses. *Divers. Distrib.* **2010**, *16*, 65-76.
35. Adami, C. Information theory in molecular biology. *Phys. Life Rev.* **2004**, *1*, 3-22.
36. Gatenby, R.A.; Frieden, B.R. Information theory in living systems, methods, applications, and challenges. *Bull. Mathemat. Biol.* **2007**, *69*, 635-657.

37. Polley, H.W.; Wilsey, B.J.; Derner, J.D.; Johnson, H.B.; Sanabria, J. Early-successional plants regulate grassland productivity and species composition: a removal experiment. Oikos **2006**, *113*, 287-295.
38. Boyero, L.; Pearson, R.G.; Bastian, M. How biological diversity influences ecosystem function: a test with a tropical stream detritivore guild. *Ecol. Res*. **2007**, *22*, 551-558.
39. Westermeier, R.L.; Brawn, J.D.; Simpson, S.A.; Esker, T.L.; Jansen, R.W.; Walk, J.W.; Kershner, E.L.; Bouzat, J.L.; Paige, K.N. Tracking the long-term decline and recovery of an isolated population. *Science* **1998**, *282*, 1695-1698.
40. Stirling, A. A general framework for analysing diversity in science, technology and society. *J. Roy. Soc. Interface.* **2007**, *4*, 707-719.
41. Keylock, C.J. Simpson diversity and the Shannon /wiener index as special cases of a generalized entropy. *Oikos* **2005**, *109*, 203-207.
42. Hill, M.O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **1973**, *54*, 427-432.
43. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys*. **1988**, *52*, 479-487.
44. A similar equation to Tsallis [43] can be found in Havrda, M.; Charvat, F. Quantification method of classification processes: concept of structural a-entropy. *Kybernetik* **1967**, *3*, 30-35, cited in [61].
45. Jost, L. The relationship between eveness and diversity. *Diversity* **2010**, *2*, 207-232.
46. Gosselin, F. An assessment of the dependence of evenness indices on species richness. *J. Theor. Biol.* **2006**, *242*, 591-597.
47. Wright, S. The genetical structure of populations. *Ann. Eugen*. **1951**, *16*, 323-354.
48. Heller, R.; Siegismund, H. Relationship between three measures of genetic differentiation G_{ST} D_{EST} and G'_{ST}: how wrong have we been? *Mol. Ecol.* **2009**, *18*, 2080-2083.
49. Ricotta, C.; Marignani, M. Computing β-diversity with Rao's Quadratic Entropy: a Change of Perspective. *Divers. Distrib.* **2007**, *13*, 237-241.
50. SPSS. http://www.spss.com/ (Accessed on 8 July 2010)
51. Log-Linear Analysis for an AxBxC Contingency Table. http://faculty.vassar.edu/lowry/abc.html (Accessed on 8 July 2010)
52. Pavoine, S.; Bonsall, M.B. Biological diversity: distinct distributions can lead to the maximization of Rao's quadratic entropy. *Theoret. Pop. Biol.* **2009**, *75*, 153-163.
53. Welsh, A. Mathematics, Australian National University (in prep, pers comm.).
54. Everitt, B.S. *The Analysis of Contingency Tables*; CRC Press: Boca Raton, FL, USA, 1992.
55. Chao, A.; Jost, L.; Chiang, S.C.; Jiang, Y.H.; Chazdon, R.L. A two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* **2008**, *64*, 1178-1186.
56. Sarrazin, C.; Bruckner, M.; Herrmann, E; Ruster, B.; Bruch, K.; Roth, W.K.; Zeuzem, S. Quasispecies heterogeneity of the carboxy-terminal part of the *E2* gene including the *PePHD* and sensitivity of Hepatitis C virus 1b isolates to antiviral therapy. *Virology* **2001**, *289*, 150-163.

57. De Bello, F.; Thuiller, W.; Leps, J.; Choler, P.; Clement, J.C.; Macek, P.; Sebastia, M.T.; Lavorel, S. Partitioning of functional diversity reveals the scale and extent of trait convergence and divergence. *J. Veget. Sci.* **2009**, *20*, 475-486.
58. Excoffier, L.; Smouse, P.E.; Quattro, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **1992**, *131*, 479-491.
59. Page, R.D.M.; Holmes, E.C. *Molecular Evolution: a Phylogenetic Approach*; Blackwell Science: Malden, MA, USA, 1998.
60. Hardy, O.J.; Jost, L. Interpreting and estimating measures of community phylogenetic structuring. *J. Ecology.* **2008**, *96*, 849-852.
61. Ricotta, C.; Szeidl, L. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoret. Pop. Biol.* **2006**, *70*, 237-243.
62. Allen,B.; Kon, M.; Bar-Yam, Y. A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *Amer. Natur.* **2009**, *174*, 236-243.
63. Guiaşu, S. Weighted entropy. *Rep. Mathl. Phys.* **1971**, *2*, 165-171.
64. Pavoine, S.; Love, M.S.; Bonsall, M.B. Hierarchical partitioning of evolutionary and ecological patterns in the organization of phylogenetically-structured species assemblages: application to rockfish (genus: *Sebastes*) in the Southern California Bight. *Ecol. Lett*. **2009**, *12*, 898-908.
65. Crozier, R.H. Preserving the information content of species: genetic diversity, phylogeny and conservation worth. *Ann. Rev Ecol Syst.* **1997**, *28*, 243-268.
66. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **1992**, *61*, 1-10.
67. Cadotte, M.W.; Davies, J.; Regetz, J.; Kembel, S.W.; Cleland, E.; Oakley, T.H. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol. Lett*. **2010**, *13*, 96-105.
68. Hohl, M.; Ragan, M.A. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst. Biol.* **2007**, *56*, 206-221.
69. Daskalakis, C.; Roch, S. Alignment-Free Phylogenetic Reconstruction. In *Research in Computational Molecular Biology*, *14th Annual International Conference*, *RECOMB 2010*, *Lisbon*, *Portugal*; Berger, B., Ed.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 123-137.
70. Otu, H.H.; Sayood, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* **2003**, *19*, 2122-2130.
71. Ané, C.; Sanderson, M.J. Missing the forest for the trees: phylogenetic compression and Its implications for inferring complex evolutionary histories. *Syst. Biol*. **2005**, *54*, 146-157.
72. Utilsky, I.; Burstein, D.; Tuller, T.; Chor, B. The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol*. **2006**, *13*, 336-350.
73. Wu, G.A.; Jun, S.R.; Sims, G.E.; Kim, S.H. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc. Natl. Acad. Sci. USA*. **2009**, *106*, 12826-12831.
74. Cantor, R. M.; Lange, K.; Sinsheimer, J.S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application amer. *J. Hum. Genet*. **2010**, *86*, 6-22.
75. Kimura, M.; Ohta, T. Distribution of allele frequencies in a finite population under stepwise production of neutral alleles. *Proc. Natl. Acad. Sci. USA*. **1975**, *72*, 2761-2764.

76. Etienne, R.S.; Olff, H. A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.* **2004**, *7*, 170-175.
77. Vellend, M. Species diversity and genetic diversity: parallel processes and correlated patterns. *Amer. Natur.* **2005**, *166*, 199-215.
78. Halliburton, R. *Introduction to Population Genetics*; Pearson Education: Upper Saddle River, NJ, USA, 2004.
79. Rossetto, M.; Kooyman, R.; Sherwin, W.B.; Jones, R. Dispersal limitations, rather than bottlenecks or habitat specificity, can restrict the distribution of rare and endemic rainforest trees. *Amer. J. Bot.* **2008**, *95*, 321-329.
80. Ryman, N.; Leimar, O. G_{ST} is still a useful measure of differentiation: a comment on Jost's D. *Mol. Ecol.* **2009**, *18*, 2084-2087.
81. Kosman, E.; Leonard, K.J. Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. *New Phytol.* **2007**, *174*, 683-696.
82. Wienberg, R. Point: hypotheses first. *Nature* **2010** *464*, 678.
83. Golub, T. Counterpoint: data first. *Nature* **2010**, *464*, 679.
84. Kang G.; Yue W.; Zhang J.; Huebner M.; Zhang H.; Ruan Y.; Lu T.; Ling Y.; Zuo Y.; Zhang, D. Two-stage designs to identify the effects of SNP combinations on complex diseases. *J. Hum. Genet.* **2008**, *53*, 739-746.
85. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **2009**, *10*, 393-404.
86. Dong, C.; Chu, X.; Wang, Y.; Wang, Y.; Jin, L.; Shi, T.; Huang, W.; Li, Y. Exploration of gene-gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* **2008**, *16*, 229-235.
87. Chanda, P.; Sucheston, L.; Zhang, A.; Brazeau, D.; Freudenheim, J.L.; Ambrosone, C.; Ramanathan, M. Ambience: A novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* **2008**, *180*, 1191-1210.
88. Chanda, P.; Sucheston, L.; Liu, S.; Zhang, A.; Ramanathan, M. Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits. *BMC Genom.* **2009**, *10*, 509
89. Volkov, I.; Banavara, J.R.; Hubbell, S.P.; Maritane, A. Inferring species interactions in tropical forests. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 13854-13859.
90. Greenspan, G.; Geiger D. Model-based inference of haplotype block variation. *J. Computat. Biol.* **2004**, *11*, 495-506.
91. Cui, Y.; Kang, G.; Sun, K.; Qian, M.; Romero, R.; Fu, W. Gene-Centric genomewide association study via entropy. *Genetics* **2008**, *179*, 637-650.
92. Laurie, C.C.; Nickerson, D.A.; Anderson, A.D.; Weir, B.S.; Livingston, R.J.; Dean, M.D.; Smith, K.; Schadt, E.E.; Nachman, M.W. Linkage disequilibrium in wild mice. *PLoS Genet.* **2007**, *3*, e144.
93. Padhukasaharsam, B.; Wall, J.D.; Marjoram, P.; Nordborg, M. Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics* **2006**, *174*, 1517-1528.
94. Siegmund, D.; Yakir, B. *The Statistics of Gene Mapping*; Springer: New York, NY, USA, 2007.

95. Stephan, W.; Song, Y.S.; Langley, C.H. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **2006**, *172*, 2647-2663.
96. Voight, B.F.; Kudaravalli, S.; Wen, X.; Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS. Biol.* **2006**, *4*, e72.
97. McVean, G. The structure of linkage disequilibrium around a selective sweep. *Genetics* **2007**, *175*, 1395-1406.
98. Clarke, C.A.; Sheppard, P.M. Further studies on the genetics of the mimetic butterfly *Papilio memnon* L. *Phil. Trans. Roy. Soc. London. B*, *Biol. Sci*. **1971**, *263*, 35-70.
99. Slatkin, M.; Excoffier, L. Maximum likelihood estimation of haplotype frequencies in a diploid population. *Mol. Biol. Evol*. **1995**, *12*, 921-927.
100. Slatkin, M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **2008**, *9*, 477-485.
101. Weir, B.S.; Hill, W.G.; Cardon, L.R. Allelic association patterns for a dense SNP map. *Genet. Epidemiol*. **2004**, *27*, 442-450.
102. Sved, J.A. Linkage disequilibrium and its expectation in human populations. *Twin Res. Hum. Genet.* **2008**, *12*, 35-43.
103. Zhang, L.; Liu, J.; Deng, H.W. A multilocus linkage disequilibrium measure based on mutual information theory and its applications. *Genetica* **2009**, *137*, 355-364.
104. Hampe, J.; Schreiber, S.; Krawczak, M. Entropy-based SNP selection for genetic association studies. *Hum. Genet.* **2003**, *114*, 36-43.
105. Reyes-Valdes, M.H.; Williams, C.G. An entropy-based measure of founder informativeness. *Genet. Res.* **2005**, *85*, 81-88.
106. Frankham, R.; Ballou, J.; Briscoe, D.A. *Introduction to Conservation Genetics*; Cambridge University Press: Cambridge, UK, 2010.
107. Madsen, T.; Stille, B. Inbreeding depression in an isolated population of Adders. *Vipera brevis Biol. Conserv*. **1996**, *75*, 113-118.
108. Hedrick, P.W. Gene flow and genetic restoration: the Florida panther as a case study. *Conserv. Biol.* **1995**, *9*, 996-1007.
109. Sommer, S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* **2005**, *2*, 16.
110. Saakian, D.B.; Fontanari, J.F. Evolutionary dynamics on rugged fitness landscapes: exact dynamics and information theoretical aspects. *Phys. Rev. E* **2009**, *80*, 041903.
111. Sella, G.; Hirsh, A.E. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci.* **2005**, *102*, 9541-9546.
112. Iwasa, Y. Free fitness that always increases in evolution. *J. Theor. Biol.* **1988**, *135*, 265-281.
113. Barton, N.H.; De Vladar, H.P. Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics* **2009**, *181*, 997-1011.
114. Mustonen, V.; Lässig. M. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4248-4253.

115. Schwanz, L.E.; Proulx, S.R. Mutual information reveals variation in temperature-dependent sex determination in response to environmental fluctuation, lifespan and selection. *Proc. R. Soc. B.* **2008**, *275*, 2441-2448.
116. Ohta, T. Linkage disequilibrium due to random genetic drift in subdivided populations. *Proc. Natl. Acad. Sci. USA.* **1982**, *79*, 1940-1944.
117. Black, W.C.IV.; Krafsur, E.S. A fortran program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theoret. Appl. Genet.* **1985**, *70*, 491-496.
118. Smouse, P.E. Likelihood analysis of recombinational disequilibrium in multiple locus gametic frequencies. *Genetics* **1974**, *76*, 557-565.
119. Smouse, P.E. Likelihood analysis of geographic variation in allelic frequencies. II. The logit model and an extension to multiple loci. *Theoret. Appl. Genet.* **1974**, *45*, 52-58.
120. von Kodolitsch, Y.; Berger, J.; Rogan, P.K. Predicting severity of haemophilia A and B splicing mutations by information analysis. *Haemophilia* **2006**, *12*, 258-262.
121. Pielou, E.C. *Mathematical Ecology*, 2nd ed.; Wiley: New York, NY, USA, 1977.
122. Gregorius, H.R. A diversity-independent measure of evenness. *Amer. Natur.* **1990**, *136*, 701-711.
123. Neilsen, R.; Tarpy, D.R.; Reeve, H.K. Estimating effective paternity number in social insects and the effective number of alleles in a population. *Mol. Ecol.* **2003**, *12*, 3157-3164.
124. Schneider, T.D.; Stormo, G.D.; Gold, L.; Ehrenfreucht, A. Information content of binding sites on nucleotide sequences. *J. Molec. Biol.* **1986**, *188*, 415-431.
125. Chao, A.; Shen, T.J. Nonparametric estimators of Shannon's index of diversity when there are unseen species in sample. *Envir. Ecol. Statistics.* **2003**, *10*, 429-443.
126. Shervais, S.; Zwick, M. Ordering genetic algorithm genomes with reconstructability analysis. *Intl. J. Gen. Syst.* **2003**, *32*, 491-502.
127. Zeeberg, B. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genom. Res.* **2002**, *12*, 944-955.
128. Wang, Y.M.; Ray, S.C.; Laeyendecker, O.; Ticehurst, J.R.; Thomas, D.L. Assessment of hepatitis C virus sequence complexity by electrophoretic mobilities of both single- and double-stranded DNAs. *J. Clin. Microbiol.* **1998**, *36*, 2982-2989.
129. Schneider, T.D. Evolution of biological information. *Nucl. Acids Res.* **2000**, *28*, 2794-2799.
130. Sanchez, R.; Grau, R. A genetic code Boolean structure. II. The Genetic Information system as a Boolean Information System. *Bull. Math. Biol.* **2005**, *67*, 1017-1029.
131. Gilchrist, M.A.; Shah, P.; Zaretzki, R. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* **2009**, *183*, 1493-1505.
132. Loewenstern, D.; Yianilos, P. Significantly lower entropy estimates for natural DNA sequences. *J. Comput. Biol.* **1999**, *6*, 125-142.
133. Díaz, J.; Alvarez-Buylla, E.R. Information flow during gene activation by signaling molecules: ethylene transduction in Arabidopsis cells as a study system. *BMC Syst. Biol.* **2009**, *3*, 48.
134. Adami, C.; Hintze, A. Evolution of complex modular biological networks. *PLoS Comput. Biol.* **2008**, *4*, e23.

135. Lezon, T.R.; Banavar, J.R.; Cieplak, M.; Maritan, A.; Fedoroff, N.V. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci.* **2006**, *103*, 19033-19038.
136. Peakall, R.; Smouse, P.E. GenALEx 6: Genetic analysis in excel. Population genetic software for teaching and research. *Mol. Ecol. N.* **2006**, *6*, 288-295.
137. msBayes. http://msbayes.sourceforge.net/ (Accessed on 8 July 2010)
138. Microsatellite Analyzer. http://i122server.vu-wien.ac.at/MSA/MSA_download.html (Accessed on 8 July 2010)
139. Refoufi, A.; Esnault, M.A. Population genetic diversity in the polyploid complex of wheatgrasses using isoenzyme and RAPD data. *Biol. Plant.* **2008**, *52*, 543-547.
140. Markwith, S.H.; Stewart, D.J.; and Dyer, J.L. TETRASAT: A program for the population analysis of allotetraploid microsatellite data. *Mol. Ecol. N.* **2006**, *6*, 586-589.
141. Pielou, E.C. The measurement of diversity in different types of biological collections. *J. Theoret. Biol.* **1966**, *13*, 131-144.
142. Horn, H.S. Measurement of "overlap" in comparative ecological studies. *Amer. Natur.* **1966**, *100*, 419-424.

Appendix: Glossary

Allele: alternative versions of the DNA sequence at a locus; see p_i.

Balance: "a function of the pattern of apportionment of elements across categories". This is based on what is called p_i in this article.

Balancing selection: various types of selection which tend to maintain variants. Also called stabilizing selection in multilocus cases.

Base: a component of DNA, also (somewhat loosely) called a nucleotide. There are four possible bases, A C T and G. The sequence of bases in DNA spells out the code. In the case of portions that code for amino acids, the code is read in triplets of bases.

δ_{ij} : represents some estimate of difference between the types (e.g., difference of morphology of species, or number of non-shared bases between alleles). See disparity.

Diploid: This is when cells contain two genomes, one from each parent individual, so that each gene might be represented by two different alleles in the one individual. Much of the information in humans is diploid. Where there is only one genome, as for the Y-chromosome or mitochondrial or chloroplast DNA, this is called haploidy (not mono- or uni-ploidy, as one might expect!). Polyploidy is when there are more than two genomes in each cell.

Differential entropy *h*: A continuous version of Shannon entropy. See Equation 17.

DNA: carries the genetic code. Composed of bases. Some parts of the code are in triplets.

Disparity: the manner and degree in which the elements of a group (e.g., the different allele types) may be distinguished. See δ_{ij} .

Directional selection: selection which eliminates one genetic variant in favour of another. Also called positive selection when the focus is on the favoured variant, or negative when the focus is on the disadvantageous variant.

Disruptive or divergent selection: when phenotypically intermediate genotypes are at a disadvantage.

Dominant: in heterozygotes, where the two different alleles from each parent are not the same type, sometimes it is only possible to detect the phenotypic effect of one allele–the dominant allele. The other allele is said to be recessive.

Drift: random processes in transmission of genes from one generation to the next.

Effective number of alleles (entropic) n_{eS}: n_{eS} is the number of equi-frequent alleles that would be required to provide the same ${}^{S}H$ value as the actual sample–see Equation 4. This is the entropic analogue of n_{eH}.

Effective number of alleles (heterozygosity) n_{eH}: n_{eH} is the number of equi-frequent alleles that would be needed to give the same heterozygosity as the actual sample–see Equation 2. Also see n_{eS}.

Effective population size: see N_e.

Epistasis: interaction between the effects of two different loci, in production of the phenotype.

Evenness: a transform of one of the diversity indices (usually Shannon's) to make explicit the departure from the most diverse case: equal numbers of each type of allele–see Equations 5 and 6.

Fitness: a function of the survival and reproduction of carriers of a certain genotype. Genotypes with higher fitness will tend to become more numerous over the generations. See also "selection".

Gametic disequilibrium: see "linkage".

Gene: this word is used variously to mean locus or allele. In the present review, it is restricted to meaning a protein-coding locus. The word should probably be abandoned, due to its sloppy use.

Genome: a complete set of genetic information, coded as base sequence of DNA. Some of this code is in triplets which each specify an amino acid in a protein. Other parts of the genome have other functions, such as regulating the expression of parts of the genome.

Genotype: the alleles contained in an individual for one or more loci.

${}^{S}H$: see Shannon entropy.

H_e : see heterozygosity.

h: see differential entropy.

Haploid: see diploid.

Haplotype: a block of DNA containing multiple SNPs, and coding for one or more genes and their regulatory regions. Haplotypes are an example of genetic linkage.

Haplotype diversity: see Heterozygosity.

Heterozygote: an individual whose genotype has one copy of each of two different alleles, at a diploid locus.

Heterozygosity, H_e: the chance of drawing two different alleles at random (with replacement) from a population: see Equation 1. Note that in this review, I do not also discuss the observed heterozygosity–the actual occurrence of heterozygous individuals. See supplement of Sherwin *et al*. 06 [1] for more discussion of this, as well as its information- theoretic applications. Heterozygosity is also called "Simpson index" when applied to species in ecological communities, "Haplotype Diversity" when it is the chance of drawing two different haplotypes at random, or "Nucleotide Diversity" when it is the chance of drawing two different nucleotides at random.

Homozygote: an individual whose genotype has two copies of the same allele at a diploid locus.

Indel: an insertion or deletion which appears in one sequence when compared to another sequence. These occur naturally during evolution of DNA. During reconstruction of phylogenies, the size and relative positions of indels must be estimated in a trade-off with the number of mismatched bases at other positions [59].

Infinite alleles model (IAM): see mutation.

Information gain: see Kullback-Liebler.

Kullback-Liebler divergence: for a given set of observed proportions of different types p_i, this is a comparison of the entropy based on an underlying distribution which really is given by p_i, versus the entropy if the underlying proportions follow some other distribution, q_i. Also called relative entropy or information gain.

$$D_{KL}(p\|q) = -\sum_i p_i \log q_i - \left[-\sum_i p_i \log p_i\right]$$

Linkage: Linkage of two different genetic loci in the genome is when the inheritance of allelic variants at one locus is not statistically independent of alleles at another. Apparent linkage between two loci is called "linkage disequilibrium" or a more correct term "gametic disequilibrium" which recognizes that apparent linkage can be due to causes other than actual physical linkage.

Linkage disequilibrium: see "linkage".

Locus: a position in the genome. Sometimes restricted to a protein-coding region, other times applied to any region of DNA at a fixed location in the genome, such as a SNP.

μ: see "mutation".

Mutation: a change to the genetic code. Note that this is best called a change, not an error–all current codes, advantageous and deleterious, were derived via multiple mutations. Various different types of mutation occur. Two contrasting types that are commonly modeled are infinite allele model (IAM), and stepwise mutation model (SMM). In IAM, every mutation makes a novel allele, which is a reasonable approximation of the evolution of a coding region made up of thousands of bases, each with four alternatives, A C G T, and a per-base mutation rate such as $\mu = 10^{-9}$ per generation. SMM or similar is seen in repetitive regions such as CACACACACACA, where repeats (CA) are added or subtracted, so that alleles of the same length are re-created regularly.

Mutual information : For two variables, the mutual information between them is the reduction in uncertainty of the level of one variable, when there is information about the level of the other variable. Or, roughly stated, this is the ability of one type of information to enlighten us about another. For example, if two populations have no shared genetic variants, then knowing the genotype of an individual would give a perfectly accurate guide to the individual's population membership, so there is said to be high mutual information between the genes and the population membership. Conversely, if the two populations have exactly the same arrays of genetic variants, then knowledge of the genes gives no indication of population membership, so mutual information is zero. See Equation 9.

N_e : effective population size: This depends not only upon actual population size, but also on any other factor that alters the rate at which random processes affect genetic quantities such as the heterozygosity H_e [78]

n_{eH} , n_{eS} : see effective number of alleles.

Negative selection: see directional selection.

Nucleotide: see base.

Nucleotide diversity: see heterozygosity.

p_i : the proportion of entities of type *i* in some group (e.g., numbers of different allelic variants encountered in a population, or numbers of different species encountered in an ecological community). See balance.

Phenotype: the detectable effect of genetic and environmental information. This might be shape, chemistry, or colour of the organism carrying a certain genotype, or the survival and reproduction of that individual.

Phylogeny: a reconstruction of the evolutionary history of a number of separate groups, usually based only upon present-day data from those groups [59].

Polymorphism: the occurrence of more than one variant within a population, e.g., two different alleles at the same locus.

Polyploid: see diploid.

Positive selection: see directional selection.

Q, or Quadratic Entropy: a generalization of Simpson's index/Heterozygosity:

$$Q = \sum_{i=1}^{S} \sum_{j=1}^{S} \delta_{ij} p_i p_j \quad \text{(Equation 10)}$$

where δ_{ij} represents some estimate of difference between the types (e.g., difference of morphology of species, or number of non-shared bases between alleles).

r_1,r_2: the proportion of a species that is in each of two populations 1 and 2. This may sometimes also be used as the relative sizes of the samples form the two populations, when performing significance testing. Note that these symbols are not to be confused with the correlation between uniting gametes, r^2, used in linkage analysis.

Reaction norm: a measure of the interaction between genotypes and environmental conditions, in production of phenotypes.

Recessive: see dominant.

Recombination: this occurs when two haplotypes from different genomes break and rejoin to make new combinations of the alleles at the different loci, ie new haplotypes.

Relative entropy: see Kullback-Liebler.

Richness: see variety.

RNA: a molecule similar to DNA, e.g., messenger RNA which carries the DNA code to the cell to be converted to an amino acid sequence in a protein.

S: see variety.

s: see selection.

${}^{S}H$ see Shannon's diversity or entropy.

Selection: the consequence of fitness differences. Genotypes with higher fitness will tend to become more numerous over the generations. See also directional, balancing and disruptive selection. Relative fitness of different genotypes is often expressed by selection coefficents *s*, where one genotype is arbitrarily assigned maximum fitness of 1, and other genotypes are given fitnesses reduced by a selection coefficient s, so their fitness is 1-*s* ($0 \le s \le 1$).

Shannon's diversity or entropy: ${}^{S}H = -\sum_{i=1}^{S} p_i \log p_i$ (Equation 3)

Simpson index: see heterozygosity.

Single-nucleotide-polymorphism: Where there is variation at one base position, this is called a "single-nucleotide-polymorphism" or SNP. Thus the alleles of a SNP locus are alternative bases A C G or T.

SNP: see "single-nucleotide-polymorphism".

Splicing: after RNA code is transcribed from the DNA code, often portions of the code are removed, between two splice sites, before the code is used to direct the production of proteins.

Stepwise mutation (SMM): see mutation.

Variety: "the number of categories into which system elements can be apportioned". Also called "richness" in biology, e.g., the number of different allelic types or the number of different species, termed *S* in this article.