How to measure genetic heterogeneity

Ryo Yamada

Professor, Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-0815, JAPAN
E-mail: ryamada@genome.med.kyoto-u.ac.jp

Abstract. Genetic information of organisms is coded as a string of four letters, A, T, G and C, a sequence in macromolecules called deoxyribonucleic acid (DNA). DNA sequence offers blueprint of organisms and its heterogeneity determines identity and variation of species. The quantitation of this genetic heterogeneity is fundamental to understand biology. We compared previously-reported three measures, covariance matrix expression of list of loci (pair-wise $r^2$), the most popular index in genetics, and its multi-dimensional form, $\Psi$, and entropy-based index, $\epsilon$. Thereafter we proposed two methods so that we could handle the diplotypic heterogeneity and quantitate the conditions where the number of DNA sequence samples is much smaller than the number of possible variants.

1. Introduction

Parents hand down the blueprint of life to children. The blueprint consists of macromolecules, deoxyribonucleic acid (DNA), coding information in a string of four letters, A, T, G and C. The sequence is copied before the inheritance with two types of modifications, mutations and recombinations. These modifications are the origin of genetic heterogeneity among biological population. The accumulation of the modifications is believed to result in speciation. The genetic heterogeneity increases the phenotypic heterogeneity and it is believed that the phenotypic heterogeneity enables species to survive against environmental changes because a fraction of the population has higher chance to be resistant to the changes.

The evolution and speciation is one of the most important topics in genetics.[1] The study to investigate relation between genetic and phenotypic heterogeneity among individuals of a species, particularly of homo sapience, is in the spotlight of medical genetics.[2] The former and the latter are the studies of inter-species and intra-species genetic heterogeneity, respectively. In this paper, I characterize the conventional measures of DNA sequence heterogeneity and introduce new measures thereafter. I discuss these measures from mathematical and genetic standpoints to prompt further studies to develop better methods to quantitate genetic heterogeneity.

2. Genetics

Although species take various reproduction strategies, I focus on one reproduction system, sexual reproduction of diploid organisms, to which homo sapience belongs. DNA in such system is classified into autosomal and sex chromosomes. The majority of DNA is autosomal. Every individual has two copies of autosomal chromosomes regardless of sex and the number of copy of sex chromosomes depends on sex. In this paper I handle autosomal heterogeneity.
See Figure 1 describing the reproduction system with mutations and reproductions. An individual G2 has a pair of DNA sequences, one of which comes from its mother (G1-1) and the other from its father (G1-2). When G2 gives DNA sequence to its offspring (G3), the maternal and paternal DNA sequences of G2 are copied and a segmental mosaic of them is produced. Therefore DNA sequence G2 gives to its offspring is different from G2’s two sequences. The phenomenon creating mosaic is called recombination. Besides recombinations, mutations also make descending DNA sequence different from its parental sequences. Mutations are the stochastic events to exchange letters of DNA sequence among A, T, G and C.

Mutations create genetic variants and recombinations exchanges combinations of variants. Because mutations are not uncommon events, DNA sequence of a species would get saturated with so many variations, if all variants persist in a population, which is not the case. Because the selection between maternal and paternal sequences to make descending mosaic is stochastic and because population size is finite, fraction of each variant fluctuates over time and sometimes variants extinct in the population. This phenomenon is called genetic drift. Actually the majority of new-born variants do not survive more than a couple of generations due to genetic drift.

3. Space of DNA sequence
Although length of DNA molecules among species vary a lot, I set its length at a fixed value, for example $L = 3 \times 10^9$, that is an approximate length of human genome. As a chemical macromolecule, $4^L$ types of sequence are possible. This is the space of DNA sequence with the
length of $L$. Because genes code functionally important information, such as particular amino acid sequence and regulation rules of their expression, a part of $L$ loci have to be a particular letter. Therefore the space is biologically restricted to $4^l < 4^L$ where $l < L$. As far as the studies of genome of *homo sapience* have revealed, almost all the variable sites are dichotomous, or have only two of four letters. Therefore $2^l$ rather than $4^l$ is sufficient as a model. In this model the sequence strings can be expressed with 0 and 1 (Figure 2).

For a region with $l$ loci with two alleles, there are $2^l$ potential variants. The $2^l$ variants can be represented as vertices of $l$-dimensional hypercube in discrete mathematics and geometry [3]. The hypercube can be drawn in a lattice as shown in Figure 3. The heterogeneity of DNA sequence is determined by frequency vector of $2^l$ elements.

![Figure 2. Example of 0 and 1 expression of sequence strings. The left on the right is fraction.](image)

![Figure 3. $l$-dimensional hypercube or lattice. $l$-dimensional hypercubes are drawn as graphs in 2-dimensional space. The number of nodes is $2^l$ and each node has $l$ edges. The power set of a set $S$ with $l$ elements consists of $2^l$ subsets including the empty set and $S$ itself. Inclusion relations among the elements of power set is drawn as a lattice, in which the subsets with the same number of elements are arranged on the same level. The number of horizontal levels is $l + 1$. The top level has only one node that is the empty set and the bottom level has also only one node, representing $S$ itself.](image)

4. Notation
- Sequence is a string of two letters with length $l$.
- Locus is an individual position in sequence: $L_i$, $i = 1, 2, ..., l$.
- Set of loci is a set with $l$ elements: $S = \{L_1, L_2, ..., L_l\}$.
- Subsets of loci is a set of loci with 0 to $l$ elements: $sS_j = \{L_{j1}, L_{j2}...\}$.
- $N_{sS_j}$ represents the number of elements of $sS_j$.
- Power set of $S$ has $2^l$ elements: $PS = \{sS_1, sS_2, ..., sS_{2^l}\}$.
- $N_{tS_j} = 2^{N_{sS_j}}$ types are defined for $sS_j$: $T_{sS_j} = \{t_{sS_j}(1), t_{sS_j}(2), ..., t_{sS_j}(N_{tS_j})\}$.
- Fraction is defined for types: $F_{sS_j} = \{f_{sS_j}(1), f_{sS_j}(2), ..., f_{sS_j}(N_{tS_j})\}$. $\sum_{k=1}^{N_{tS_j}} f_{sS_j}(k) = 1$
• Haploid and Diploid: Diploid organisms have a pair of DNA sequences, each of which comes from one of parents. Haploid means a half set of DNA and Diploid means a set of two DNA sequences.

• $N_{dip, ss_j}$ is the number of diploid types of $ss_j$. When the diploptype of $(t_{ss_j}(x), t_{ss_j}(y))$ and the diploptype of $(t_{ss_j}(y), t_{ss_j}(x))$ are considered the same, $N_{dip, ss_j} = \binom{N_{t_{ss_j}}}{1} + \binom{N_{t_{ss_j}}}{2}$.

When they are considered discriminable, $N_{dip, ss_j} = (N_{t_{ss_j}})^2$. The former definition is useful usually but the latter is necessary for particular biological phenomena, which needs to discriminate maternal and paternal sequences.

• $T_{dip, ss_j} = \{t_{dip, ss_j}(1), t_{dip, ss_j}(2), ..., t_{dip, ss_j}(N_{dip, ss_j})\}$.

• Fraction of $T_{dip, ss_j}$ is $F_{dip, ss_j} = \{f_{dip, ss_j}(1), f_{dip, ss_j}(2), ..., f_{dip, ss_j}(N_{dip, ss_j})\}$.

5. Methods to quantify haploid sequence heterogeneity

The methods to quantify haploid sequence heterogeneity can be divided into two types. One type quantitates sequence heterogeneity with multiple parameters. The other type gives only one or a few values to describe the sequence heterogeneity as a whole.

5.1. Variance-covariance matrix and $r$ or $r^2$

The most popular method to describe sequence heterogeneity in genetic studies is similar to $l \times l$ variance-covariance matrix of $l$ loci.

Consider two subsets of one locus, $ss_{j1} = \{L_{j1}\}$ and $ss_{j2} = \{L_{j2}\}$ and their joint subset $ss_{j3} = \{L_{j1}, L_{j2}\}$. They have $F_{ss_{j1}} = \{f_{ss_{j1}}(1), f_{ss_{j1}}(2)\}$ and $F_{ss_{j2}} = \{f_{ss_{j2}}(1), f_{ss_{j2}}(2)\}$ and $F_{ss_{j3}} = \{f_{ss_{j3}}(1), f_{ss_{j3}}(2), f_{ss_{j3}}(3), f_{ss_{j3}}(4)\}$. The covariance between $L_{j1}$ and $L_{j2}$, $r_{j1,j2}$, is given as,

$$r_{j1,j2} = \frac{f_{ss_{j3}}(1)f_{ss_{j3}}(4) - f_{ss_{j3}}(2)f_{ss_{j3}}(3)}{\sqrt{f_{ss_{j1}}(1)f_{ss_{j1}}(2)f_{ss_{j2}}(1)f_{ss_{j2}}(2)}},$$

(1) where

$$f_{ss_{j1}}(1) = f_{ss_{j1}}(1) + f_{ss_{j3}}(2),$$

$$f_{ss_{j1}}(2) = f_{ss_{j3}}(3) + f_{ss_{j3}}(4),$$

$$f_{ss_{j2}}(1) = f_{ss_{j1}}(1) + f_{ss_{j3}}(3),$$

$$f_{ss_{j2}}(2) = f_{ss_{j3}}(2) + f_{ss_{j3}}(4).$$

Actually $r_{j1,j2}^2$ is used regularly in genetic studies and is called linkage disequilibrium index and it ranges from $-1$ to $1$. $\left(\frac{l}{2}\right)$ pair-wise indices are plotted to display the heterogeneity visually as shown in Figure 4. $\chi^2$ test has a good reason to be used in genetic studies as follow.

Assume a study to find genetic locus that is associated with a phenotype and assume $\chi^2$ test is performed to test the association between markers and the phenotype. When a marker itself is functionally related with the phenotype, $\chi^2$ value for the marker is expected as the strength of the association indicates. Let $\chi^2(L_i)$ denote the expected value. In many cases of the studies to identify such locus, markers close to the true locus, that are indirectly associated with the phenotype, are tested. The $\chi^2$ value of those markers are expected to be

$$\chi^2(L_j) = r_{i,j}^2\chi^2(L_i),$$

where $L_i$ is not functionally associated with the phenotype and $L_i$ and $L_j$ are mutually dependent with $r_{i,j}$. This relation is important because regular genetic studies to identify phenotype-causing loci have to make a part of loci represent all loci without acceptable loss of power by a limited number of loci. This method summarizes the genetic heterogeneity which requires $2^l - 1$ parameters to be fully described, with $\left(\frac{l}{2}\right)$ parameters instead.
Figure 4. \( \binom{l}{j} \) values are plotted in a triangle with color-intensity scale. The values of neighboring loci are plotted on the top horizontal line and the values of loci in longer distance are plotted in the lower area.

Figure 5. The upper half is the triangle of pair-wise \( r^2 \) in gray scale. The lower half is the triangle of subsets with elements in tandem. The cells on the central and horizontal line represent the pairs of loci side-by-side. The cells above the central line represent the pairs of loci separated by one locus in-between. The cells above next are for the pairs separated by two loci. The cells below the central line represent the trios of loci in tandem. The cells below next for the quartets of loci in tandem and so on.

5.2. Power set LD index and their partial display, \( \Psi \)

This method extracts information of \( 2^l \) types with \( 2^l - 1 \) parameters. This method uses the fact that the power set of a set of \( l \) elements consists of \( 2^l \) subsets as,

\[
\sum_{i=0}^{l} \binom{l}{i} = 2^l.
\]

This equation is demonstrated by the fact that the elements of power set can be drawn in a lattice as shown in Figure 3.

\( \Psi \) consists of \( 2^l - 1 \) parameters, each of which is the value for an element of power set except for the empty set as \( \Psi = \{ \psi_{sS_j} \} \) and they are defined as [6]

\[
\psi_{sS_j} = \sum_{k=1}^{N_{sS_j}} v_{sS_j}(k) \times f_{sS_j}(k),
\]

where

\[
v_{sS_j}(k) = \prod_{x=1}^{N_{sS_j}} \sigma_{sS_j}(k)x = 1 \text{ or } 1,
\]
when the x-th locus of $t_sS_k (k)$ is one arbitrarily determined type

$$\sigma_{sS_j}(k)_x = 1,$$

and when the x-th locus of $t_sS_k (k)$ is the other

$$\sigma_{sS_j}(k)_x = -1.$$

Another simple way to express fraction distribution of $2^l$ types is to give one parameter for each type except the last type. Compared to this way, the parameters of $\Psi$ are hierarchically structured and the parameters for smaller subsets tend to represent wider picture of its structure than ones for larger subsets. Although this method expresses all information of sequence heterogeneity, the major drawback is that the number of parameters is too big to display all the information. Therefore it might be the palliative answer to chose $2 \times \binom{l}{2}$ parameters among $2^l - 1$, which are displayed in a square and one half of them is identical to the triangle plot of $r^2$ and the other half represents subsets whose elements are in tandem. Figure 5 is the drawing for sequence fraction in Figure 2. [6]

5.3. Entropy-based measures of LD or heterogeneity, $\epsilon$

This measure describes sequence heterogeneity with one value. It is based on the information theory. The definition of entropy $H$ is

$$H = - \sum_{i=1}^{2^l} f_i \ln(f_i) \ (f_i \neq 0).$$

(3)

$H$ takes the minimum when only one type exists and the value is

$$H_{\text{min}} = 0.$$    

(4)

$H$ takes the maximum when all types exist with identical fraction and the value is

$$H_{\text{max}} = - \sum_{i=1}^{2^l} \frac{1}{2^l} \times \ln\left(\frac{1}{2^l}\right) = l \times \ln2.$$  

(5)

Nothnagel et al. defined $\epsilon$ using $H$. They supposed that $l$ loci had variants and the fraction of variants were given. Actually when fraction of all types is the expected value under the assumption of independence among all loci,

$$H = \sum_{k=i}^l H_i,$$

where $H_i$ is entropy of $L_i$. On the other hand when only two types exist, entropy of all individual loci are identical as $H_i$ and

$$H = H_i.$$

Under the condition where fractions of individual loci were given, $\epsilon$ was standardized so that it gave the minimum value, 0, for the state where fraction of all types is the expected value under the condition all loci are mutually independent and it gave the maximum value, 1, for the state where only two out of $2^l$ types exist. [7] Two major benefits of this measure is:

- One value represents the whole.
- It ranges in the fixed range, from 0 to 1, regardless of $l$.
- Its calculation only uses fraction of existing types. Because the number of existing types is limited, this feature is useful.
5.4. An extension of $r^2$, $r^l$

$\epsilon$ has very favorable features as an index of heterogeneity as listed in the previous section. But $\epsilon$ for two loci is not identical with $r$ or $r^2$. Because $r$ or $r^2$ are closely related with measures for statistical evaluation such as correlation and test of independence, an index that was extended from $r$ or $r^2$ and that had the characters of $\epsilon$ listed above would be useful. Following descriptions will develop such index, $r^l$ from $r^2$.

$r^2_{i,j}$ can be expressed as below;

$$ r^2_{i,j} = \frac{1}{4} \sum_{k=1}^{4} (fs_{S_{j3}}(k) - E(fs_{S_{j3}}(k)))^2 E(fs_{S_{j3}}(k)),$$

or

$$ r^2_{i,j} = \frac{1}{4} \sum_{k=1}^{4} \frac{(fs_{S_{j3}}(k))^2}{E(fs_{S_{j3}}(k))} - 1,$$

(6)

where

$$ E(fs_{S_{j3}}(1)) = fs_{S_{j3}}(1) \ast fs_{S_{j3}}(1),$$

$$ E(fs_{S_{j3}}(2)) = fs_{S_{j3}}(1) \ast fs_{S_{j3}}(2),$$

$$ E(fs_{S_{j3}}(3)) = fs_{S_{j3}}(2) \ast fs_{S_{j3}}(1),$$

$$ E(fs_{S_{j3}}(4)) = fs_{S_{j3}}(2) \ast fs_{S_{j3}}(2).$$

Of note $r^2$ is $\chi^2$ statistics of test of independence between two loci standardized for sample size. Let $\chi^2(st)$ denote $\chi^2$ statistics for independence standardized for samples size. Then

$$ r^2 = \chi^2(st).$$

For $sS_i$ with loci more than 2, independence of fraction under the condition where fraction of variants of individual loci is given, can be quantitated with

$$ \chi^2(st) = \sum_{k=1}^{N_{tS_i}} \frac{(fs_{S_i}(k) - E(fs_{S_i}(k)))^2}{E(fs_{S_i}(k))},$$

or

$$ \chi^2(st) = \sum_{k=1}^{N_{tS_i}} \frac{(fs_{S_i}(k))^2}{E(fs_{S_i}(k))} - 1.$$  

This is a simple extension of $r^2$ for 2 loci. However this $\chi^2(st)$ ranges from 0 to 1 when $l = 2$ but it ranges from 0 to an unfixed value when $l > 2$. This unfixed range makes $\chi^2(st)$ unfavorable as an index. Therefore I defined $r^l$ by modifying $\chi^2(st)$ so that it ranges from 0 to 1 regardless of $l$ as

$$ r^l = \sum_{k=1}^{N_{tS_i}} \frac{(fs_{S_i}(k))^{l+1}}{E(fs_{S_i}(k))^{l+1}} - 1.$$  

(7)

See Appendix for further description. Because $l = 2$, equations (6) and (7) are identical, $r^l$ is an extension of $r^2$ for two loci and is a modification of $\chi^2$ statistics for independence overall and satisfies the favorable features of $\epsilon$. The further statistical properties of $r^l$ have yet to be investigated.
6. Further requirements for diploids and many loci

6.1. From haploids and diploids

The methods described above, \( r, r^2, \Psi, \epsilon \), and \( r^l \), were defined to quantitate frequency data of haplotypes but not of diploids, even when the phases of diploid genotypes are known. It is sufficient to compare sequences of various species, because one sequence representing each species is treated. However it is necessary to quantitate genetic heterogeneity among diploid individuals who possess a pair of sequences. In genetics when individuals randomly mate, combination of sequences in diploid individuals is expected independent. When combination of sequences in individuals is not independent, non-random mating is usually indicated. The randomly-mating condition is genetically called "in Hardy-Weinberg equilibrium (HWE)" and the condition that is most remote from "HWE" is the condition where the population is completely subdivided into two homogeneous subpopulations ("complete subdivision"). Therefore genetic heterogeneity of diploids is heterogeneity of haploid sequence and mating. Although \( r \) or \( \Psi \) is not easily extended to handle diploids data, \( \epsilon \) and \( r^l \) can be. Entropy which is the base of \( \epsilon \) is calculable by fraction of diploid types.

\[
H\text{(diploid)} = - \sum_{k=1}^{N_{dip,sS_j}} f_{dip,sS_j}(k) \times \ln(f_{dip,sS_j}(k)).
\]  

(8)

It would be reasonable to define the "most" and "least" heterogeneous states of diploids as below. When haploid sequence entropy is the maximum and when diploids are in HWE, the state is "most" heterogeneous. When haploid sequence entropy is the minimum and when diploids are completely subdivided, the state is "least" heterogeneous. The values of \( H \) for both conditions are expressed as,

\[
H\text{(diploid, max)} = 2 \sum_{i=1}^{l} H_i,
\]

\[
H\text{(diploid, min)} = H_i.
\]

It would be appropriate to calculate \( \epsilon \) for diploids by standardizing entropy of diploid fraction data with the maximum and minimum entropy values for diploids as given above.

In order to make \( r^l\text{(diploids)} \) range from 0 to 1, the exponent is changed from \( l^{-1} \) to \( \frac{1}{2l-1} \) as

\[
r^l\text{(diploids)} = \sum_{k=1}^{N_{dip,sS_j}} \left( \frac{f_{dip,sS_j}(k)}{E(f_{dip,sS_j}(k))} \right)^{\frac{1}{2l-1}+1} - 1.
\]  

(9)

The calculation of \( H \) and \( r^l \) for diploids are affected by \( N_{dip,sS_j} \), which can vary as described in Notation section. Briefly when maternal and paternal sequences are discriminated, information of data is more than the condition without the discrimination. Details on this issue are beyond this paper.

6.2. Sparsity and "Biological" heterogeneity measure

\( l \) is quite large, such as \( 3 \times 10^9 \) for \textit{homo sapien}ce, therefore all individuals are genetically different even though human population is more than \( 6 \times 10^9 \) worldwide. \( sSs \)s with much less loci are still able to distinguish all individuals. When it is the case, \( f_{sS_j}x \) or \( f_{dip,sS_j}y \) is \( \frac{1}{N} \) or 0 where \( N \) is sample size. Because both \( H \) and \( r^l \) are functions of \( f_{sS_j}x \) or \( f_{dip,sS_j}y \), fixed values of the fraction are not appropriate. Actually \( H = \ln(N) \) regardless of existing types.

Suppose any one haploid or diploid individual is different from all others. Then the population can be considered as a discrete distribution in \( l \)-dimensional hypercube or a lattice graph with \( 2^l \) nodes, in which only \( N \) nodes are occupied by one individual and all other nodes are vacant,
where the space is sparsely occupied. This sparse distribution would be the subject of geometry or graph theory.

Here reconsider from biological and genetic standpoint whether it is appropriate to locate \(2^l\) types in \(l\)-dimensional hypercube or its lattice graph counterpart. When DNA sequences among various species are compared, distance between any two sequences is accumulation of mutations. Therefore distance in the space or in the lattice corresponds to biological distance. However when DNA sequences among a population of same species are compared, their relations are determined both by mutations and recombinations. This means the distribution in \(l\)-dimensional hypercube or the lattice might not be appropriate for sequences in the same species. Two types of distance can be defined for sequences in the same species. One distance is defined between any two sequences as the number of mutations, that changes one sequence to the other. The other distance is defined between a sequence (offspring sequence) and a pair of sequences (parent sequences) as the number of recombinations, which creates the offspring sequence from parent sequences (Figure 6).

**Figure 6.** Distance by recombination. Sequence pair of 1 run creates a sequence of \(k\) runs with \(k-1\) recombinations. The recombination distance of this offspring sequence from the parents as \(k-1\). A given pair of sequences that are different at \(l\) locus has \(2 \times \left(\frac{l}{k}\right)\) offsprings that are \(k\) recombination distance from them \((0 \leq k \leq l - 1)\).

**Figure 7.** Each line represents an order of samples, in which the samples are removed from the population, and the decrease of heterogeneous loci is drawn. The inner plot in the lower left is one of 100 lines and the slashed area was calculated to extract representing values from 100 lines.

This indicates that straightforward mathematical approaches based on geometry or graph-theory might not be relevant for combination of mutations and recombinations. Because possession of some amount of mutations makes a species survive through environmental stresses, it is beneficial to have variants in the way where genetic drift will not easily lose the variants excessively. When some fraction of population is removed from the whole, some loci become homogeneous in the remnant. Figure 7 shows the relation between the fraction of population extinction and the fraction of loci remaining heterogeneous, for a population that consists of 19
samples, each of which corresponds to one of the sequences shown in Figure 2. One hundred permutations of 19 samples were randomly selected and samples were withdrawn from the total one by one in the order of each permutation and the fraction of withdrawn samples and the fraction of heterogeneous loci in the remnant were calculated and plotted. When the fraction of extinction was 0, all loci remained heterogeneous and when only one individual remained, all loci are homogeneous. Therefore all lines in Figure 7 started at the top left and ended at the bottom right. The orders, in which individuals were removed from the population, determined the pathway from the top left down to the bottom right. Based on this idea, I randomly obtained the extinction lines in Monte-Carlo methods and quantitated the lines of multiple trials with average and variance of area above the extinction lines. Some conditions have similar average value of the extinction lines but different variance value. Therefore at least two values from the extinction simulations, average and variance, will be necessary to quantitate their heterogeneity. The example shown in Figure 7 was one of haploid data but it is simple and easy to have its diploid counterpart.

7. Brief summary
In this paper, I described three methods previously reported and proposed two new methods. Two of five, variance-covariance matrix, and Ψ, describe heterogeneity of haploids but not diploids with multiple values and they are not for sparse data. Another two of five, entropy-based ϵ and χ²-related rl, describe data with one value and are applicable to diploid data, but not to sparse data. The last method is simulation-based and two descriptive values from simulation results seem to represent haploid and diploid heterogeneity and it is applicable to sparse data. (Table 1) Only a few simple rules, mutations, recombinations and genetic drifts, can produce very complex discrete data and further studies are necessary to develop better methods to quantitate the complexity.

Table 1. Control sequences to describe lines and symbols in figure captions.

| Methods                              | No. parameters | Sparse data | Diploid |
|--------------------------------------|----------------|-------------|---------|
| Variance-covariance matrix of r or r² | \( \binom{l}{2} \) | +           | -       |
| Power set system, Ψ                 | \( 2^l - 1 \)  | +           | -       |
| Entropy-based ϵ                     | 1              | -           | +       |
| Standardized χ², \( r^l \)         | 1              | -           | +       |
| Extinction-resistance               | 2              | +           | +       |

8. Appendix
When there are \( l \) loci, the condition where all loci have the same allele frequency and all loci are mutually in absolute linkage disequilibrium, is the most deviate state from the equilibrium. Let \( p \) denote their allele frequency, then only two haplotypes exist in the population and their frequencies are \( p \) and \( 1 - p \). Their expected frequencies in linkage equilibrium are \( p^l \) and \( (1 - p)^l \), respectively.

\[
r^l = \frac{p^{l-1} + 1}{(p^l) \frac{1}{1-l}} + \frac{(1 - p)^{l-1} + 1}{((1 - p)^l) \frac{1}{1-l}} - 1
\]

\[
r^l = p + (1 - p) - 1
\]

\[
r^l = 1.
\]
References

[1] Reznick D N and Ricklefs R E 2009 Darwin’s bridge between microevolution and macroevolution Nature 457 837–42
[2] Altshuler D, Daly M J and Lander E S 2008 Genetic Mapping in Human Disease Science 322 881–8
[3] Henk M, Richter-Gebert J and Ziegler G M 2004 Basic properties of convex polytopes Handbook of Discrete and Computational Geometry (2nd ed) ed J E Goodman and J O’Rourke (Boca Raton, FL: Chapman & Hall/CRC) chapter 16
[4] Devrin B and Risch N 1995 A comparison of linkage disequilibrium measures for fine-scale mapping Genomics 29 311–322
[5] Barrett J C, Fry B, Maller J and Daly M J 2005 Haploview: analysis and visualization of LD and haplotype maps Bioinformatics 21 263–5
[6] Yamada R and Matsuda F 2007 A novel method to express SNP-based genetic heterogeneity, Ψ, and its use to measure linkage disequilibrium for multiple SNPs, D(g), and to estimate absolute maximum of haplotype frequency Genet. Epidemiol. 31 709–26
[7] Nothnagel M, Furst R and Rohde K 2002 Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks Hum. Hered. 54 186–98