Geographic Barrier Increases and Elongates Extensive Linkage Disequilibrium

Joon Ho Kang
Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Dated: May 20, 2013)

Quantitative study of gene interaction can reveal gene function, mutation and their spatial and time propagation. Due to complicate phenotype-genotype correlation, inferring epistasis from mere phenotypic observation misleads our investigation path. In this study, we propose a computational model of geographically separated genotypes evolving under the constant competition between epistatic interaction and recombination. We demonstrated the existence of two different selection regimes, and quantitatively investigated its transitional behavior using Linkage Disequilibrium (LD). Via increasing the genetic drift by geographically separating genotypes, we successfully observed the elongation of genotype selection regime (higher critical recombination rate, r_c, for transition into allele selection regime) and sharpening of the transition. Our results propose that even a small, discrete physical separation can possibly have an enormous impact on the genotypic distribution.

Thorough understanding of interactions among numerous genes can greatly nourish our knowledge of nature, because interaction often reveals gene function, mutation and their spatial and time propagation [1]. Extensive theoretical modeling and numerous experiments on epistasis in fact prove the strong desire of many scientists across a broad range of field. Recent studies have in fact shown that epistasis accounts for observed variety of phenotypes by a significant fraction [2]. Other studies in C.elegans or plants also support the evidence of widespread epistasis and further investigated the effect of outcrossing events in selection dynamics [3-4]. Epistatic interaction in recombining population plays a major role in selection dynamics, since positive interaction between loci will likely to proliferate its genotype across the generation but recombination will partly randomize the genotypic interaction by breaking up co-inherited loci pair. However, exponentially growing number of phenotypes and its opaque correlation with the genotype, previous experimental approach is not adequate for quantifying the competition between epistasis and recombination.

The first quantitative approach to epistasis was done by a number of theorists. Kimura, using two loci haploid organism, first showed that linked gene systems rapidly settle down to a Quasi Linkage Equilibrium (QLE) under selection dynamics [5]. Subsequent studies in different regime collectively produced a wide-known term, Linkage Disequilibrium (LD), which has enormous significance in providing quantitative information on the strength of gene interaction. LD has an importance in evolutionary biology and human genetics because of its ability to infer past population history and genotypic distribution. Yet, much is known about the multi-loci population due to statistical difficulties encountered when genotypic distribution and interaction are correlated with wide range of phenotypic expression and also because of limited theoretical knowledge. Thus, recent studies on multi-loci genetic population were mostly done by intensive simulation. Neher and Shraiman showed that average linkage or correlation between allele frequencies in QLE theory, which approximates the linkage in weak epistasis/high recombination region, makes a sharp transition into clonal competition region with abnormal increase in linkage between alleles in fit genotype [6]. This remarkable transition held for a wide range of population (10^3 to 10^5), number of loci (20-100) with simple random epistasis per genotype assigned at the beginning of propagation. However these studies have assumed the complete randomness in mating, which is unrealistic given the enormous range of spatial distribution of individuals.

Here we investigate the effect of non-random mating on the transition between genotype selection to allele selection regime by simulating the time course of genotypic distribution. Perhaps, more applicable to evolutionary dynamics will be non-random mating, since physical distance and regional isolation partly disrupts the random

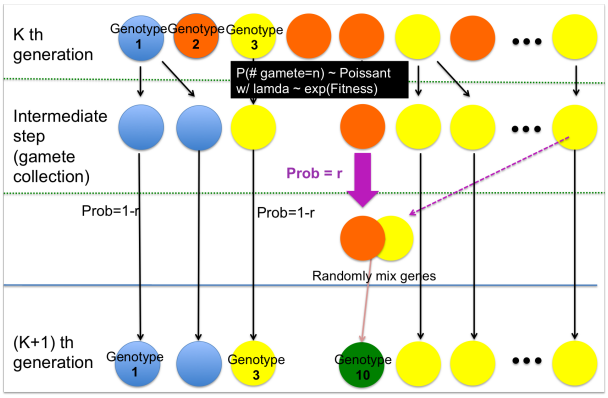

FIG. 1. Schematics of simulation algorithm. For each generation, individuals' genotype are represented by decimal value of binary genotype. Each individual produce a number of copies drawn from a poissant distribution with normalized parameter exp(Fitness). Then, each progeny has a probability $1-r$ to asexually propagate its genotype or a probability r to randomly blend their genes by mating. By collecting both asexually propagated genotypes (no mate) and re-assorted genotype (mate), genotypic distribution of the next generation is complete.

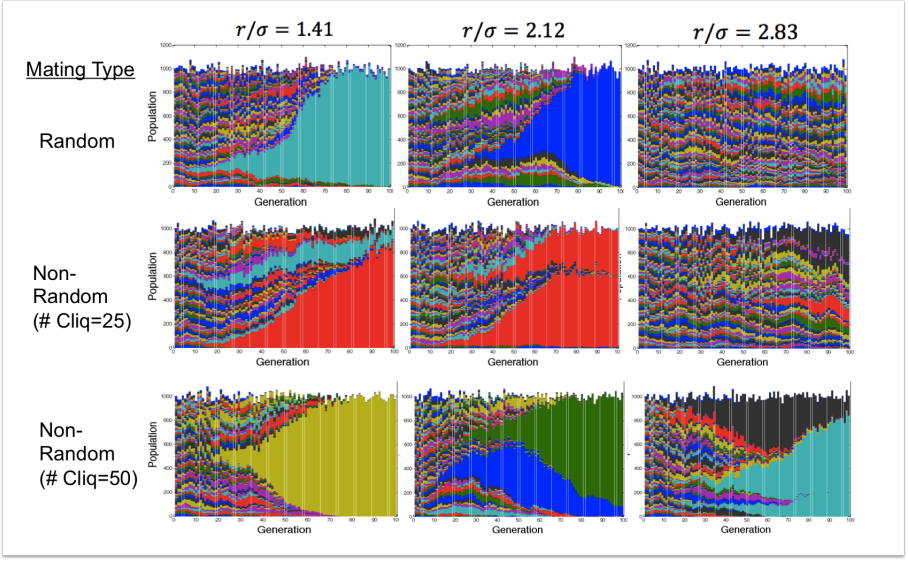

FIG. 2. Simulated evolution of genotypic distribution. ($N_pop = 10^3$, $V_i = 0.005$, $V_a = 0$) Along the row, outcrossing rate r ranges from 0.1, 0.15, 0.2. After gamete collection, same genotypes were grouped together and stacked on top of each other. Therefore, the relative height of each color denotes the relative frequency of each genotype. The proliferation of fit genotype, represented by a dominance of one color stripe, is notable for all three mating scheme in the low outcrossing rate ($\frac{r}{\sigma} = 1.41$). In relatively high regime of outcrossing, population evolved with intense geographic barrier still presents the genotype selection whereas in complete random mating scheme, only partial proliferation is readily available.

probability of finding its mate. How does the transition between two regime affected by the intensity of physical barrier? To answer this, we have done extensive simulation on the linkage of multi-loci population in multi-region separated space. In this study, we partially incorporate Neher and Shermans Random Epistasis model (RE model) in order to justify the validity of our model.

For each genotype g, the fitness, which is a measure of one's ability to pass on its complete or partial copy to next generation is given by,

$$F(g) = f \sum_{i}^{L} s_i + \xi(g) \tag{1}$$

where $f = \sqrt{\frac{V_A}{L}}$, and V_a from a normal distribution with mean 0 and variance V_I.

For the time-course simulation of genotypic distribution, we have used the binary model for describing each individuals genotype; for each locus i, either binary variable, $s_i = +1$ or $s_i = -1$, determines its additive contribution to the total fitness of genotype. $\xi(g)$ in the above equation denotes for random epistatic fitness, which is drawn from normal distribution with mean 0 and variance V_I depending on the genotype, yet fixed over time. Since epistatic fitness was assigned in the beginning of simulation depending on the genotype, it is not heritable even if the genotype of a progeny slightly differs from that of parents.

Once the fitness of each genotype is specified, we have accordingly used poissant distribution with mean parameter proportional to an exponential of relative fitness., ie, $exp(F - F_{ave})$, to create a pool of gamete in each generation. Then, by a given range of recombination probability, two gametes were allowed to randomly re-assort their genes. In order to mimic the geographic barrier in mating, we assigned a probability weighted accordingly to the geographic distance between the individuals. Instead of assigning geographic location to each individual, for simplicity, we have introduced a certain number of cliques, and subdivided the population into each clique randomly at the beginning of simulated time course. In this way, the mating probability is simply weighted by the difference in clique index of individuals.

We now examine the actual results of extensive computer simulation. We initialized simulations in a random, genetically diverse state with L=8 - 20 and N=10^3-10^4 and observed the evolution of genotypic distribution depending on the wide range of outcrossing rate and number of geographic barriers in non-random mating. Often, the genotypic distribution can be qualitatively understood visually, since Extraordinary linkage provides global pattern different from that produced by mere sampling noise. Fig.2 qualitatively visualize the effect of introducing geographic barrier on transition from genotype selection (broad band type) and allele selection (messy dot type). First row of figures visually shows the breakdown of Genotype selection in relatively high recombination regime, and the transition to allele selection regime

is notable by the thin and dispersed pattern of genotype. This is because in low recombination regime, fit genotypes with positive interaction between alleles will likely to proliferate and propagate, whereas in the relatively high recombination regime, alleles are constantly reshuffled to disrupt the positive interaction and thus propagation of its genotype. Interestingly, increasing the intensity of geographic barrier by enlarging the number of cliques for individuals to be distributed, transition to quasi linkage equilibrium was delayed as the last column of the figure shows the growing pattern of bands or stripes across the generation. This infers that the critical recombination rate for a transition to allele selection regime (messy dot pattern) is increasing as the geographic isolation between individuals is intensified.

In order to probe the transition behavior more quantitatively, we have computed the genome-wide LD. Conventional LD denotes the deviation of probability for observing a pair of alleles from the expectation based on each individual probability. To indicate the deviation as a correlation coefficient, we have used another common way of quantifying LD given by

$$LD_{genome} = \sum_{i<j} \psi_{ij}^2 \tag{2}$$

where $\psi_{ij}^2 = \frac{(<s_i s_j> - <s_i><s_j>)^2}{v_i \bar{v}_i v_j \bar{v}_j}$ should be a time invariant LD per locus pair even if individual allele frequencies v_i and v_j does change over the time course [Kimura]. Although LD should be time invariant In order to measure average LD over large population evolving in time, for couple generation in the beginning of the simulated time course, LD greatly fluctuates due to the erratic nature of gamete production and mating. Thus, in order to choose some stable LD, we have picked the average LD_{genome} to be the LD value when allelic entropy given by $S_A = -\sum_i [v_i ln(v_i) + \bar{v}_i ln(\bar{v}_i)]$ reduces to 0.7 of its initial value. Figure 3(A) shows the average LD per locus as a function of outcrossing rate for several values of geographic barrier intensity. ($L = 20, N_{pop} = 10^3$). The average LD versus outcrossing rate directly presents two interesting sequential changes: 1) In general, population under more geographic pressure (higher number of cliques) shows higher LD for all range of outcrossing rates, which more or less agrees with the visual pattern of linkage as we discussed above in Figure 2. Astonishingly, the LD values after fixation is also higher in population evolved in larger or more geographic separation; 2) r_c changed accordingly with the intensity of geographic separation. For population evolving under high geographic pressure seems to have elongated genotype selection regime than that under low geographic separation.

It is noteworthy to discuss above two phenomenon by comparing with the case of higher population size, Above two phenomenon is more pronounce in higher population as expected from previous studies, increasing the population size both sharpens and delays (higher r_c) the transition. At first glance, the result we have obtained above seems to be challenging the previous studies, since readily mating population is in fact reduced to N_{pop}/N_C for population evolving under geographic isolation. Since breakdown of QLE approximates $r_c \simeq \sqrt{2V_I ln(rN\tau)}$, and τ does not intensely vary under different mating condition [Not shown. Supplementary Figure 1 and 2], reduction of effective population for random mating seems to be lowering r_c. However, in terms of the severity of reshuffling fit genotypes, population evolving in presence of geographic pressure will have fewer chances for its geno-

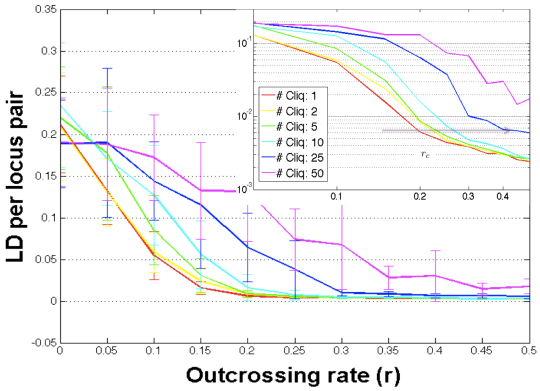

FIG. 3. Linkage Disequilibrium (LD) of population evolving in different number of geographic isolation. Data for L=20, $N_{pop} = 10^3$ averaged over 20realizations. (Inset) LD as a function of outcrossing rate in log-log plot. Notice the sequential increase of r_c as the number of cliques. $V_A = 0.1\sigma^2$, $V_I = 0.9\sigma^2$, $\sigma^2 = 0.005$

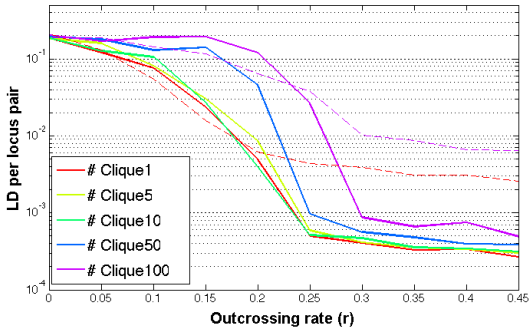

FIG. 4. Linkage Disequilibrium (LD) of population under different geographic pressure on mating. Same condition as in Fig.3, but for population size of 10^4. Similar to the case of population 1000, higher geographic stress on mating sharpens the transition, increases critical recombination rate and LD after fixation. Dashed lines are those from population 1000. The scale of transition and LD after fixation greatly differs by population size

type to be completely rendered. In other words, effective N stays about the same whereas effective epistatic variance, V_I greatly increases due to small sampling out of large population. This comes much clearer when examining the LD after fixation ($r > r_c$). Genetic drift is mostly responsible for the LD after fixation, because the effect is the same as taking a small sample from large population [7]. As shown in the inset of Figure 4, geographic separation increases the LD after fixation, opposed to increasing the population size alone, which has the exact opposite effect as introducing geographic separation.

To conclude, we have shown that relative strength of positive gene interaction to gene reshuffling determines the genotypic distribution to be in distinct regime (genotype/allele selection regime), and non-random gene shuffling changes the transitional behavior between the two regime: Intense geographic separation causes critical recombination rate to increase via genetic drift, which also accounts for the higher LD after fixation of population evolving in severe geographic isolation. The quantitative study of evolution of spatially segregated population would serve as a key for investigating subglobal or local selection and genetics.

I thank Prof. Leonid Mirny and Prof. Mehran Kardar for introducing such interesting topic and Anton Goloborov for helpful comments on the simulation planning.

[1] N.H.Barton et al., Nat Rev Genet 2, 11 (2005).
[2] R.B.Brem et al., Nature 436, 701 (2005).
[3] E.S.Dolgin et al., Evolution 61, 1339 (2007)
[4] M.Parker, Evolution 46, 837 (1992)
[5] Kimura, Genetics 52, 875 (1992)
[6] R.A. Neher et al., PNAS 106, 6866 (2009)
[7] M. Slatkin, Nat Rev Genet 9, 477 (2008)

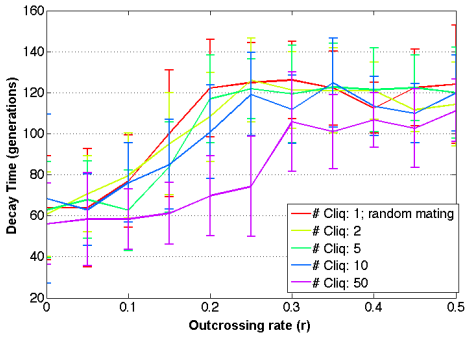

FIG. 5. Supplementary Figure. Decay time, the number of generations for allelic entropy to drop below 0.7 of its initial value, is plotted as a function of outcrossing rate. For all mating scheme, decay time were in reasonable range, showed similar trends. shown data are averaged over 20realizations and error bar denotes the standard deviation for each outcrossing rate. $N_{pop} = 10^3$, $V_A = 0.1\sigma^2$, $V_I = 0.9\sigma^2$, $\sigma^2 = 0.005$