

January 01, 2010

Bose-Einstein distribution, condensation transition and multiple stationary states in multiloci evolution of diploid populations

Ginestra Bianconi
Northeastern University

Olaf Rotzschke
Agency for Science, Technology and Research, Singapore

Recommended Citation

Bianconi, Ginestra and Rotzschke, Olaf, "Bose-Einstein distribution, condensation transition and multiple stationary states in multiloci evolution of diploid populations" (2010). *Physics Faculty Publications*. Paper 43. <http://hdl.handle.net/2047/d20000396>

This work is available open access, hosted by Northeastern University.

Bose-Einstein distribution, condensation transition and multiple stationary states in multiloci evolution of diploid populations

Ginestra Bianconi¹ and Olaf Rotzschke²

¹*Physics Department, Northeastern University, Boston, Massachusetts 02115 USA,*

²*Singapore Immunology Network (SIgN), Biomedical Sciences Institutes (BMSI), Agency for Science, Technology and Research (A*STAR), IMMUNOS, 138648, Singapore, Singapore*

The mapping between genotype and phenotype is encoded in the complex web of epistatic interaction between genetic loci. In this rugged fitness landscape, recombination processes, which tend to increase variation in the population, compete with selection processes that tend to reduce genetic variation. Here we show that the Bose-Einstein distribution describes the multiple stationary states of a diploid population under this multi-loci evolutionary dynamics. Moreover, the evolutionary process might undergo an interesting condensation phase transition in the universality class of a Bose-Einstein condensation when a finite fraction of pairs of linked loci, is fixed into given allelic states. Below this phase transition the genetic variation within a species is significantly reduced and only maintained by the remaining polymorphic loci.

PACS numbers: 89.75.Hc, 87.23.Kg

I. INTRODUCTION

The deep relation between evolutionary theory and statistical mechanics has been fascinating most of the scientists working in the field. Historically, Fisher compared his fundamental theorem of natural selection to the second law of thermodynamics [1] and Kimura, referring to the stochasticity of the evolutionary process, has compared the genetic theory of evolution to the theory of gases [2]. More recently, further relations between evolutionary theory and statistical mechanics have been identified [3, 4]. Indeed the relation between *chance and necessity* [5] in evolutionary theory, i.e. the trade-off between stochastic processes and selection, has the potential to be fully described in statistical mechanics terms. Interestingly, the relation between evolutionary theory and quantum statistical mechanics is emerging from a series of independent works [6–13] where it has been highlighted that a class of phase transitions occurring in evolution of haploid population can be described in terms of a Bose-Einstein condensation. These transitions are called in the biological literature quasispecies transitions and represent collective change in the populations when mutations compete with natural selection. A quasispecies transition that can be mapped to a Bose-Einstein condensation, occurs in asexual populations when the mutation rate is below a critical value where a finite fraction of the population is localized in sequence space around a given genotype. In this paper we show that the Bose-Einstein distribution and condensation transition in the Bose-Einstein universality class occur in the evolutionary theory of diploid populations when recombination competes with selection. Therefore we propose that Bose-Einstein statistic is the emergent statistics both in evolution of haploid and diploid populations when there is a competition between processes enhancing genetic variation in the population (i.e. mutations or recombination process) and selection processes (which tend to decrease

genetic variation in the population).

The genomic revolution started with the publication of the entire sequence of the human genome [14, 15] has made possible the complete analysis of genome variations encoded in single nucleotide polymorphisms (SNPs). The complete set of SNPs of an individual characterizes together with the copy number variations what is unique about an individual. Variations in SNPs allelic states (i.e. different nucleotide composition of the SNPs) can affect how humans develop diseases and how they respond to pathogens or drugs. It is well established [16–21], that genes are integrated in functional pathways and interact through complex biological networks. Single nucleotide polymorphisms can affect expression or function of the genes and they are correlated when gene products are part of a joint pathway. If the functional pathways constitute the phenotype, then the interaction between the complete set of SNPs and these pathways encodes for the genotype-phenotype mapping. Consequently, SNPs are interacting through the bio-molecular networks and their contribution to the fitness of an individual is encoded in complex epistatic interactions between the SNPs [22, 23].

A recent paper has signed a turn-over in the study of epistatic interactions [24]. Indeed in [24] the epistatic network between a large set of pairs of mutations in yeast, has been fully characterized. This work demonstrates the possibility of collecting data for this new fundamental type of biological network [16–21] in simple organisms, by measuring the effect on fitness of pairs of mutations in yeast. From the structural point of view, this network is both modular [21, 25] and fat tail [16, 20] as the regulatory network [21], the protein interaction network [26–28] and the metabolic network [29]. From the evolutionary point of view, this epistatic network sheds light at the genotype-phenotype relation and it reveals a functional map of the cell in which genes with highly correlated profiles delineate specific pathways. Similar networks exist also in higher organisms and, importantly, a substantial number of genes are regulated on the pop-

ulation level by the allelic states of polymorphic loci. A functional genome analysis of the signaling pathways of human thrombocytes revealed that a striking number of genes of the response cascade is under allelic control [30].

Linkage disequilibrium between SNPs [22] is a key quantity for identifying genes that are related to specific diseases. In particular, linkage disequilibrium indicates the non-random association of alleles at two or more loci (SNPs) and is widely observed through the human genome [31]. Non random mating of a population and variation of the cross-over rate and finite evolutionary times contributes to the occurrence of linkage disequilibrium in diploid populations. However, also epistatic interactions between genetic loci contributes to the observed linkage disequilibrium. There is compelling evidence that linkage disequilibrium occurs not only between genetic loci in close proximity to the chromosome but also between genetic loci at significant distance on the same chromosome or even on different chromosomes, as summarized in the scheme shown in Fig. 1. In order to explain this phenomenon it is necessary to describe the long-distance epistatic interactions between SNPs, which are not exclusively weak.

In order to develop an evolutionary theory in presence of epistatic interactions it is necessary to go beyond the well defined single locus evolution [32–35]. Nevertheless, most of the available mutiloci evolutionary theories [34] are typically limited to a few numbers of loci. A relevant exception is the recent paper of R. Neher and B. I. Shraiman [13] where the authors have studied the role of the crossover rate in an evolutionary theory of a large number of genetic loci, interacting epistatically in a network. Interestingly, they found by mean-field arguments and by numerical simulations, that the evolutionary model shows a phase transition responsible for the level of genetic variation in a population. In fact, in their evolutionary theory, for high recombination rates the population is in the "allele selection" phase in which genetic loci are only weakly associated, while for low recombination rates the population is in the "genotype selection" phase consisting of a set of competing genotypes locked in given allelic combinations.

In this paper we study the genetic variability of sexually reproducing diploid populations where free genetic recombination competes with Darwinian natural selection [36] under different strengths of the selective pressure. We consider an epistatic network of N loci (with $N \gg 1$) of a general topology, and we take the fitness of an individual as a function of the allelic states of genetic loci in this epistatic network. In order to approach the formidable task to tackle the complexity of a mutiloci evolution, we neglect mutations (that further contribute to the genetic variation in the population) and we assume free recombination and infinite population and time limit. Finally, by making an ansatz on the shape of the distribution of gametes allelic state in the population, we are able to characterize all the stationary states in linkage equilibrium (while we leave to subsequent publications

the study of solutions compatible with linkage disequilibrium). The technical improvement with respect to the previous mutiloci theories that makes our theory exactly solvable, is due to the advantages of defining the fitness function over an epistatic network and using the most recent developments of statistical mechanics [10, 37–40]. In particular in this paper we make use of a self consistent argument [10] combined with the cavity method [37–40].

The stationary states are multiple and therefore, asymptotically in time, the state of a population depends on the initial conditions. Unexpectedly, the joint frequency of allelic states of pairs of linked loci, at stationarity, is expressed in terms of a Bose-Einstein distribution. In a quantum Bose gas, the Bose-Einstein distribution describes the occupation number of energy levels. Moreover, a quantum Bose gas, below a critical temperature, might undergo a Bose-Einstein condensation transition in which a finite fraction of particles are found in the ground state. In our mutiloci evolution dynamics for diploid populations, the relation of the steady state solution with the Bose-Einstein distribution, allows us to predict a condensation transition in the Bose-Einstein universality class [41, 42]. In this evolutionary model, a pair of genetic loci is in the "ground state" when they are fixed, i.e. when they are in a given allelic state (not necessarily the same allelic state in each genetic locus) and they are not any more polymorphic. For a given value of the selective pressure, and a suitable topology of the epistatic network that allows for a condensate phase, a finite fraction of pairs of loci is fixed. Therefore a finite fraction of pairs of loci is not any more polymorphic and the number of polymorphic loci is significantly reduced. The basic mechanism behind the studied condensation of genetic loci, is a cooperative phenomenon in which, as the selection pressure increases, one locus that is fixed in a beneficial configuration, affects the other linked loci inducing them to fix in given allelic configurations, generating a macroscopic phase transition. A similar phenomenon is also known in the two-loci evolutionary dynamics and is called in the literature "genetic hitchhiking" [43]. We note here that the phase transition observed in our model should be distinguished from the "genotype selection" of [13] because the genotypes in the population maintain a significant genetic variation due to the remaining polymorphic loci which have not been fixed. Interestingly, another phase transition between a polymorphic phase and a frozen phase has been numerically observed at a critical value of the mutation rate in sexual populations [44].

The paper is divided as follows: in Sec. II we define the fitness function that drives the evolutionary dynamics in presence of a complex epistatic network; in Sec. III we define the evolutionary dynamic equation under consideration; in Sec. IV we highlight the results regarding the steady state population of the evolutionary dynamics, including all the details of the derivations in the subsequent appendixes; and finally we give the conclusions.

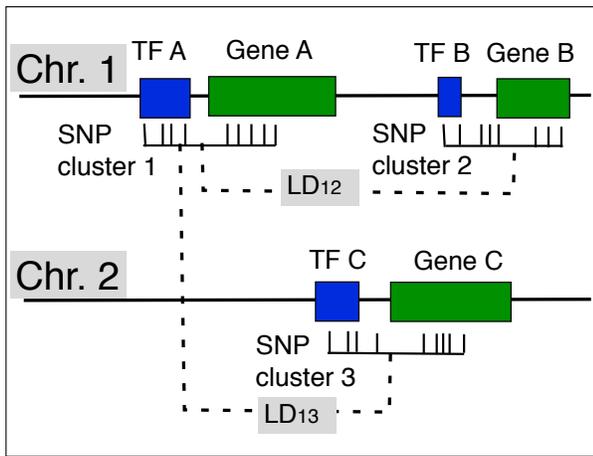


FIG. 1: (Color online) Linkage disequilibrium (*LD*) between clusters of SNPs. Single nucleotide polymorphisms are usually arranged in small clusters with all members being in complete linkage (haplotype). SNPs epistatic interactions are mediated by the transcription factors (TF) binding on regulatory regions and by the genes interacting in regulatory and signaling networks. Linkage disequilibrium between clusters of single nucleotide polymorphism (SNPs) can either appear when two clusters are in close proximity along a chromosome (as for example LD_{12} in the figure) or when the SNPs clusters are at a significant distance along the chromosome or even in two different chromosomes (as for example LD_{13} in the figure). While different recombination rates might explain part of the linkage disequilibrium for clusters of SNPs in proximity along the chromosome, only epistatic interactions can explain linkage disequilibrium for distant clusters.

II. THE FITNESS FUNCTION AND THE EPISTATIC NETWORK

Haploid cells have a single copy of each chromosome. On the contrary, diploid cells have two homologous copies of each chromosome (see Fig. 2). Usually the genome of each diploid individual is given by the pairs of chromosomes *A* and *B* of the two haploid gametes coming from the father and from the mother of the individual. Let us suppose that each gamete is identified by N loci indicated with latin letters $i = 1, 2, \dots, N$. If we indicate with x_i the allelic state at each locus i , the gamete is characterized by the sequence $\{x_i\}_{i=1,2,\dots,N}$ with $x_i = 1, 2, \dots, Q$ and Q given by the biochemistry of the DNA, i.e. $Q = 4$ indicating respectively the pair of ordered nucleotides AT (Adenine, Thymine), TA (thymine, adenine), CG (cytosine, guanine) or GC (guanine, cytosine) in the double stranded DNA. Given this description of the gametes, each individual is characterized by the sequence $\{x_i^A, x_i^B\}_{i=1,2,\dots,N}$ where $x_i^{A/B}$ indicates the allelic states in each parental gametes *A/B*. In the multiplicative non-epistatic (NE) scenario the fitness function $W^{NE}(\{x^A, x^B\})$ factorizes into contributions coming from independent single loci, i.e. allelic states in a pair of loci do not have epistatic interactions. In this

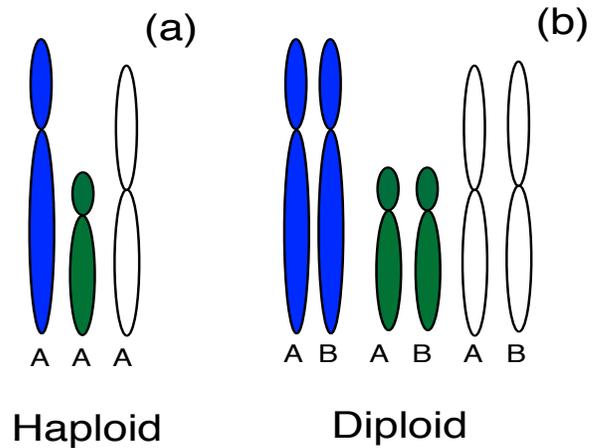


FIG. 2: (Color online) Difference between haploid and diploid cells: (a) while haploid cells have a single copy of each chromosome, diploid (b) organisms have two homologous copies of each chromosome.

traditional framework the fitness function is written as

$$W^{NE}(\{x^A, x^B\}) = \prod_i \psi_i^{NE}(x_i^A, x_i^B). \quad (1)$$

In the free recombination hypothesis, the minimal modification to this theory that is compatible with the presence of epistatic interactions is that the fitness is a function of allelic states of pairs of loci. Therefore we assume that the loci $i = 1, 2, \dots, N$ are linked in an epistatic network formed by L links. We define by ∂i the set of loci j linked with locus i in the network. The epistatic couplings between pairs of loci have a role in determining the fitness function that can be modified with respect to the single locus expression (1), according to the expression

$$W(\{x^A, x^B\}) = \prod_{\langle ij \rangle} \phi_{ij}(x_i^A, x_j^A, x_i^B, x_j^B), \quad (2)$$

where the product is extended to all genetic loci $\langle i, j \rangle$ linked in the epistatic network. Therefore, the fitness function in Eq. (2) is the first non trivial correction to (1) and includes a product of contributions coming from pairs of diploid allelic states at different loci.

We parametrize the functions $\phi_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)$ as in the following

$$\phi_{ij}(x_i^A, x_j^A, x_i^B, x_j^B) = e^{-\beta U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)} \quad (3)$$

where the parameter β indicates the selective pressure. In fact for $\beta = 0$ we recover a neutral theory while for large β small variations of the function $U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)$ yield large changes in the fitness. Furthermore the function $U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)$ has the following symmetries

$$\begin{aligned} U_{ij}(x_i, x_j, x'_i, x'_j) &= U_{ji}(x_j, x_i, x'_j, x'_i) \\ U_{ij}(x_i, x_j, x'_i, x'_j) &= U_{ij}(x'_i, x'_j, x_i, x_j). \end{aligned} \quad (4)$$

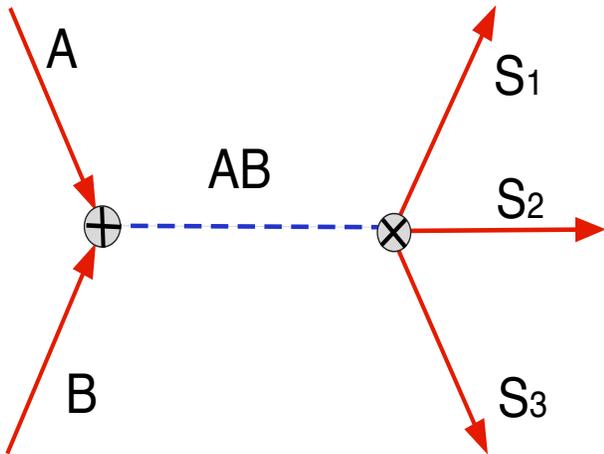


FIG. 3: (Color online) Graphical representation of the transfer of genetic information in overlapping generations called in biology the *gametic life cycle*. This diagram describes two parental gametes A, B (solid lines) that give rise to a zygote and then to an individual AB (dashed line) by a fertilization process indicated with \oplus . This individual on his or her turn generates $n = 3$ during meiosis \otimes new gametes S_1, S_2, S_3 .

with the last relation valid only if we assume perfect symmetry between the parental gametes, i.e. if we exclude to study the sex chromosomes X, Y . The fitness landscape defined in Eq. (2) corresponds to a disordered Potts Hamiltonian and therefore it is in general characterized by many local maxima.

III. EVOLUTIONARY DYNAMICS

The evolutionary dynamics of diploid populations describes the information transfer of genetic information encoded in the gamete sequences. Each individual of a diploid population is carrying the information encoded in the gametes of their parents indicated as A/B . The evolution is due to the transmission of each individual to the next generation of new gametes which are a random recombination of the information encoded in parental gametes A/B . In physical terms the process can be seen as a "scattering" process in which two gametes A/B generate a new individual (*fertilization*) and the new individual, if it reaches the reproductive state, carries the information and gives rise (by *meiosis*) to new gametes S_1, S_2, \dots, S_n with $n = 0, 1, 2, \dots$. We can visualize this process also called *gametic life cycle* by the diagram shown in Fig. 3 in which each solid line is a gamete and each dotted line is an individual. The vertices of this diagram are indicated with a sign \oplus when fertilization occurs or with a sign \otimes when meiosis occurs, i.e. a new gamete is generated by a process of recombination of the diploid genetic information. The presence of these vertices of the diagram is an exclusive characteristics of diploid organisms wherever in haploid organism the single individual of the population is transmitting the genomic informa-

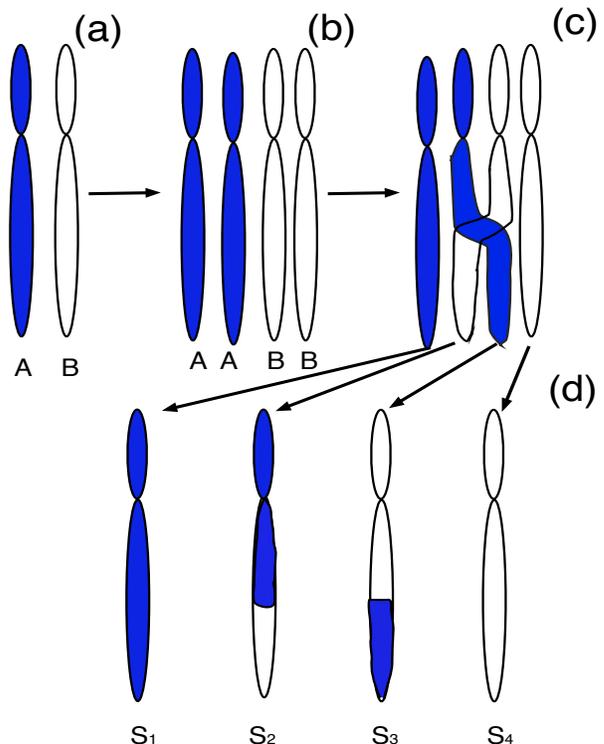


FIG. 4: (Color online) Graphical representation of meiosis. Starting from a given diploid genome (a) of an individual AB , as the outcome of replication (b), cross-over and genetic recombination (c) a set of haploid gametes S_1, S_2, \dots, S_n are formed (d). An exchange of genetic information might occur at each recombination site. In particular, every pair of genetic loci of the gamete either come from a single parent or from different parents.

tion to the next generation. The process of meiosis is a process of reduction in the genetic information of each diploid individual to generate gametes which have only half of the number of chromosomes. During meiosis (see Fig. 4) a process of recombination can occur with small probability at given locations (recombination hotspots) on the chromosome. When a recombination event occurs, homologous sites on two chromosomes can mesh with one another and may exchange genetic information. In our evolutionary dynamics we take the infinite population limit and we assume that at each genetic locus a recombination event can take place. Therefore, the probability $P(\{x\})$ that a gamete has an allelic configuration $\{x\}$, satisfies the following dynamical equation

$$\frac{\partial P(\{x\})}{\partial t} = M_{\{x\}} \left[\frac{W(\{x^A, x^B\})P(\{x^A\})P(\{x^B\})}{\langle W \rangle} \right] - P(\{x\}) \quad (5)$$

where $W(\{x^A, x^B\})$ is the fitness of the individual of diploid allelic configuration $\{x^A, x^B\}$ given by Eqs. (2) and (3) and where $\langle W \rangle$ is the average fitness

$$\langle W \rangle = \sum_{\{x^A\}, \{x^B\}} W(\{x^A, x^B\})P(\{x^A\})P(\{x^B\}). \quad (6)$$

The operator $M_{\{x\}}$ introduced in Eq. (5) indicates the free recombination of genetic material occurring when a new gamete is generated. In particular the operator $M_{\{x\}}$ is defined as the average over the probability of free recombination $\Pi(\{x\}|\{x^A, x^B\})$, i.e. the action of the operator over a generic function $f(\{x^A, x^B\})$ is given by

$$M_{\{x\}} [f(\{x^A, x^B\})] = \sum_{\{x^A\}, \{x^B\}} \Pi(\{x\}|\{x^A, x^B\}) f(\{x^A, x^B\}) \quad (7)$$

with

$$\Pi(\{x\}|\{x^A, x^B\}) = \prod_{i=1}^N \left[\frac{1}{2} \delta(x_i, x_i^A) + \frac{1}{2} \delta(x_i, x_i^B) \right]. \quad (8)$$

We note here that in this model we assume free recombination and equivalence between the parent gametes. Moreover, in order to simplify the treatment of the evolutionary model, we limit our study to evolution of diploid populations in absence of mutations.

IV. RESULTS

A. General form of steady state probability distributions

If the network is locally tree-like, the general structure of the solution to the evolutionary equation (5) is given by

$$P(\{x\}) = \sum_h \pi(h) \prod_{\langle i,j \rangle} b_{ij}^h(x_i, x_j). \quad (9)$$

where $\langle i, j \rangle$ indicates all the pairs of linked nodes present in the epistatic network.

In our model the fitness function of the type (2) and (3), for any given epistatic network, is static and bounded from above, therefore we always expect to find asymptotically in time the population in a stationary state given by the solution of to the equation

$$P(\{x\}) = M_{\{x\}} \left[\frac{W(\{x^A, x^B\}) P(\{x^A\}) P(\{x^B\})}{\langle W \rangle} \right]. \quad (10)$$

These stationary states, do not necessarily correspond to a maximum of the fitness [45]. In the case treated in this paper, in which the epistatic network is fixed and locally tree-like, we can find, for every generic fitness function of type (2) and (3), the possible stationary states of the population (see appendixes) of the type

$$P(\{x\}) = \prod_{\langle i,j \rangle} b_{ij}(x_i, x_j), \quad (11)$$

where the product is extended to all genetic loci $\langle i, j \rangle$ linked in the epistatic network. The type of solutions (11)

is a subset of the general type of solutions (9). In particular, in order for (11) to be a solution of the stationary relation Eq. (10), $P(\{x\})$ must satisfy the condition

$$P(\{x^A\}) P(\{x^B\}) = P(\{x_{-i}^A, x_i^B\}) P(\{x_{-i}^B, x_i^A\}) \quad (12)$$

where $\{x_{-i}^A\}$ indicates all the variables $\{x^A\}$ except variable $\{x_i^A\}$ and $\{x_{-i}^B\}$ indicates all the variables $\{x^B\}$ except variable $\{x_i^B\}$. It can be easily shown that these conditions, enforce linkage equilibrium between allelic states. In this paper we restrict our attention to this type of solutions and we leave to subsequent publications the study of stationary state distributions compatible with linkage disequilibrium.

The marginal frequencies $p_{ij}(x_i, x_j)$ of allelic states on pairs of linked loci (i, j) are defined as

$$p_{ij}(x'_i, x'_j) = \sum_{\{x\}} P(\{x\}) \delta(x_i, x'_i) \delta(x_j, x'_j). \quad (13)$$

B. Bose-Einstein distribution

In order to find all the stationary solutions solving Eq. (10) of the form given by (11), we used a self-consistent argument [10] combined with the cavity method [37, 38, 40]. In particular we find that in the stationary state, not necessary a maximum of the fitness function [45], the marginal frequencies $p_{ij}(x_i, x_j)$ defined in Eq. (13) are given by (see Appendix BandC)

$$p_{ij}(x_i, x_j) = \frac{1}{Z} G_{ij}(x_i, x_j) \{1 + n_B[\epsilon_{ij}(x_i, x_j)]\} \quad (14)$$

if $\epsilon_{ij}(x_i, x_j) > \mu$. The functions $\epsilon_{ij}(x_i, x_j)$, $G_{ij}(x_i, x_j)$ in Eq. (14) and the constants μ , Z can be derived from the self consistent solution of the stationary state of the evolutionary dynamics described by Eq. (5) (see Appendix B). In Eq. (14) $n_B(\epsilon)$ indicates the Bose-Einstein occupation number and μ indicates the "chemical potential" of the evolutionary dynamics. The Bose-Einstein occupation number is defined as

$$n_B[\epsilon_{ij}(x_i, x_j)] = \frac{1}{e^{\beta[\epsilon_{ij}(x_i, x_j) - \mu]} - 1}, \quad (15)$$

Equation (14) relates the joint probability of pair of linked loci with a Bose-Einstein distribution arising in the study of quantum Bose gases [41, 42]. Here the functions, $\epsilon_{ij}(x_i, x_j)$ play the role of "energy states" of this Bose-Einstein distribution. These functions are not known a priori but they are the outcome of the evolutionary dynamics. A relevant aspect of this solution is that we might find several different sets of functions $\{\epsilon_{ij}(x_i, x_j), G_{ij}(x_i, x_j)\}$ and variables Z, μ that satisfy the stationary condition of the evolutionary dynamics. These different solutions have to be identified with different possible populations of a given species. In fact, given different initial conditions the population evolving

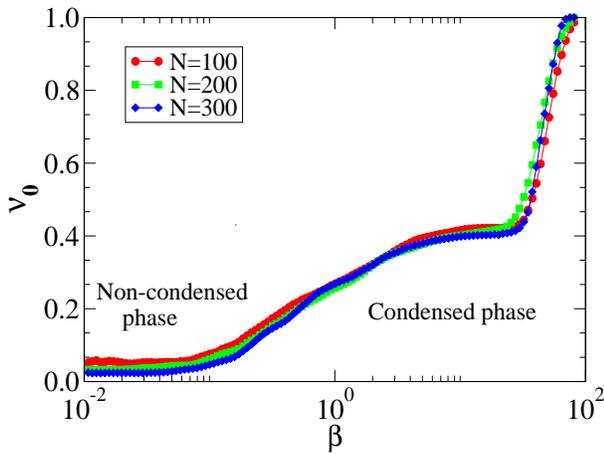


FIG. 5: (Color online) Numerical evidence for the condensation phase transition. The fraction ν_0 of fixed pairs of loci is plotted as a function of the selective pressure β . The data are averaged over 50 random fitness realizations and are shown for a epistatic network with degree distribution given by Eq. (21). The fitness function used is given by Eqs. (2) and (3) with the matrix elements $U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)$ satisfying the symmetry constraints (4) and drawn randomly from a uniform distribution in the interval (0, 1). The data are shown for a number N of genetic loci with $N = 100, 200, 300$.

according to Eq. (5) can be found, asymptotically in time, in different stationary states. According to the general multiloci evolutionary scenario [45], these steady states do not in general correspond to local maxima of the fitness landscape.

Interestingly, the marginal frequencies $p_{ij}(x_i, x_j)$ are significantly modified if a given pair of allelic configuration (x_i^*, x_j^*) reaches the minimal allowed "energy level" $\epsilon_{ij}(x_i^*, x_j^*) = \mu$. In this case we found $G_{ij}(x_i, x_j) = 0$ for every allelic state (x_i, x_j) and the pairs of linked loci (i, j) gets fixed (see Appendix D). Therefore if $\epsilon_{ij}(x_i^*, x_j^*) = \mu$, the joint probability $p_{ij}(x_i, x_j)$ is given by

$$p_{ij}(x_i, x_j) = \delta(x_i, x_i^*)\delta(x_j, x_j^*). \quad (16)$$

To be specific in the terminology used, here and in the following we assume that a pair of genetic loci is fixed if and only the joint distribution is given by (16), i.e. if and only if both genetic loci (i and j) are fixed.

C. Condensation transition

The joint distributions (14) and (16) must always satisfy the normalization constraints

$$1 = \sum_{x_i, x_j} p_{ij}(x_i, x_j) \quad (17)$$

valid for every pair of linked loci (i, j) . This set of equations plays the role of "conservation" laws for the evolutionary dynamics and determines the phase diagram of

the evolutionary dynamics. Inserting in (17) the expression (14) for $p_{ij}(x_i, x_j)$ when $\epsilon_{ij}(x_i, x_j) < \mu$ and expression (16) for $p_{ij}(x_i, x_j)$ when $\epsilon_{ij}(x_i, x_j) = \mu$ we can write the normalization conditions (17) as in the following,

$$(1 - \nu_{ij}) = \frac{1}{Z} \sum_{x_i, x_j} G_{ij}(x_i, x_j) + \frac{1}{Z} \sum_{\substack{x_i, x_j \\ \epsilon_{ij}(x_i, x_j) > \mu}} G_{ij}(x_i, x_j) n_B[\epsilon_{ij}(x_i, x_j)] \quad (18)$$

valid for every pair of linked loci (i, j) . In Eq. (18), $n_B(\epsilon)$ is the Bose-Einstein distribution defined in Eq. (15) and the matrix elements ν_{ij} are $\nu_{ij} = 1$ the pairs of linked loci (i, j) are fixed, otherwise $\nu_{ij} = 0$. Considering Eq. (18) and averaging over all pairs of linked loci we can write

$$(1 - \nu_0) = \frac{1}{Z} \int_{\epsilon > \mu} d\epsilon [1 + n_B(\epsilon)] g_\beta(\epsilon) \quad (19)$$

where $n_B(\epsilon)$ is the Bose-Einstein distribution defined in Eq. (15), and ν_0 is the fraction of pairs of linked loci that are fixed in the population. The quantities $g_\beta(\epsilon)$ and ν_0 present in Eq. (19) are given by

$$g_\beta(\epsilon) = \lim_{\Delta\epsilon \rightarrow 0} \frac{1}{\Delta\epsilon} \frac{1}{2L} \sum_{i, j \in \partial i} \sum_{x_i, x_j} G_{ij}(x_i, x_j) \chi_{\Delta\epsilon}(\epsilon_{ij}(x_i, x_j) - \epsilon) \\ \nu_0 = \frac{1}{2L} \sum_{i, j \in \partial i} \sum_{x_i, x_j} \chi_0[\epsilon_{ij}(x_i, x_j) - \mu_c]. \quad (20)$$

where $\chi_\delta(x) = 1$ if $|x| \leq \delta$ and $\chi_\delta(x) = 0$ otherwise. Depending on the form of $g_\beta(\epsilon)$ the solution of Eq. (19) might indicate the occurrence, as a function of β , of a condensation phase transition characterized by the order parameter ν_0 . For epistatic network topologies and fitness functions which display this phase transition, we can distinguish, as a function of β , between a noncondensed phase in which the fraction of fixed pairs of loci ν_0 vanishes in the thermodynamic limit, i.e. $\nu_0 \rightarrow 0$ as $N \rightarrow \infty$ and a "condensed phase" in which the fraction of fixed pairs of genetic loci is finite in the thermodynamic limit, i.e. $\nu_0 \rightarrow \bar{\nu} > 0$ as $N \rightarrow \infty$.

This condensation phase transition is in the same universality class of the Bose-Einstein condensation transition as it depends on the value of the integral of a Bose-Einstein distribution present in Eq. (19). We observe, nevertheless, that Eq. (19) differs from the equation fixing the average number of particles in a Bose gas [41, 42] because the function $g_\beta(\epsilon)$ given by Eq. (20) depends on β while the correspondent density of states in a quantum Bose gas is independent of β . Moreover in Eq. (20) there is an additional factor Z in the right hand side with respect to the correspondent equation in the quantum Bose gas.

In the noncondensed phase all genetic loci are polymorphic, on the contrary, in the "condensed phase" only a fraction of genetic loci is polymorphic. In which phase

are diploid populations usually found? If we assume that each base of the DNA is a candidate SNP, we observe that polymorphisms only occur in a finite fraction of bases. For example, in the human genome less than 1% of the bases corresponds to SNPs. Here we propose that the condensation of genetic loci due to epistatic interactions, might significantly contribute to the reduction in genetic variation within a species.

D. Numerical evidence of the condensation transition

While the results exposed in the preceding section are valid for any fitness function of type (2) and tree-like epistatic network, the actual phase diagram of the evolutionary dynamics might change depending on the topology of the network and on the detail of the fitness function. In this paragraph we show numerical evidence for the condensation of genetic loci by solving the self-consistent equations (see Appendix B) that determine $\{\epsilon_{ij}(x_i, x_j), G_{ij}(x_i, x_j)\}$ and ν_0 , for a given fitness function, starting from random initial conditions. In particular we consider a network topology that allows for long-distance epistatic interactions (see Fig. 1). We have therefore chosen to study an epistatic network with degree distribution

$$P(k) \propto k^{-\gamma}. \quad (21)$$

In Fig. 5 we show evidence for the occurrence of the condensation transition of genetic loci when the epistatic network is a random network with degree distribution given by (21) and $\gamma = 3$. In particular in Fig. 5 we have plotted the fraction ν_0 of fixed pairs of loci (averaged over several random realizations of the fitness function) as a function of the evolutionary pressure β .

The "condensed phase" is defined as the region where ν_0 is large and does not show finite size effects. Outside this region, instead, we have the "non condensed phase" where the fraction of fixed loci goes to zero in the limit of large N , i.e. $\nu_0 \rightarrow 0$ as $N \rightarrow \infty$. The fitness function used in the numerical solution reported in Fig. 5, is given by Eqs. (2) and (3) with the matrix elements $U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)$ satisfying the symmetry constraints (4) and drawn randomly from a uniform distribution in the interval $(0, 1)$. Finally, in order to reduce the time for the numerical solution of the self-consistent equations we have taken $Q = 2$.

E. Condensation transitions in evolutionary dynamics

Condensation phase transitions universally occur in evolutionary models. A pivotal condensation phase transition occurs in the quasispecies [6–8] evolutionary model of haploid populations that describes the competition between random mutations, which tend to increase the ge-

netic variation of the population, and natural selection which tends to reduce it. In the quasispecies model, when the mutation rate μ is less than a critical value μ_c , i.e. $\mu < \mu_c$, the haploid population is localized in the sequence space, and when, instead, $\mu > \mu_c$ there is no possibility to define a typical sequence in the population. A condensation transition also occurs in the "house of cards" model of Kingman [9] which describes the quasispecies model in the limit of infinite loci. Kingman characterized the condensation transition in the "house of cards" model but only recently, with the study of evolving complex systems, i.e networks [10] and ecosystems [12], and in a more elaborated model with pleiotropy [11] it was recognized that this condensation can be mapped to a Bose-Einstein condensation in a Bose gas.

Condensation phase transitions also occur in diploid populations. In [13] it was shown that the phase transition between the "allele selection" phase and the "genotype selection" phase is a condensation phase transition below which, for low recombination rates, few genotypes are selected in the population.

Here we show that a condensation transition in the Bose-Einstein universality class is also occurring in diploid populations in the presence of free genetic recombination. The novelty of this transition is that the occurrence of the Bose-Einstein statistics is not caused by mutations (as it is the case for the quasispecies and the "house of cards" models) but only by genetic recombination. Moreover, this condensation transition differs from the transition between "allelic selection" and "genotype selection" of [13] because in the condensed phase of the present multiloci evolutionary theory, the population maintains a wide variation although the number of polymorphic loci is significantly reduced. Finally, the condensation of genetic loci of diploid populations is a consequence of the non-trivial interactions of genetic loci in the epistatic network while in the quasispecies model and in the "house of cards" model the interactions between the individuals of the population are only mediated by the competition for finite resources. Therefore the condensation of genetic loci in the present evolutionary theory relates to the condensation transition in the quasispecies model [6, 8] as the condensation transition in interacting quantum Bose gases [42] relates to the condensation transition in non-interacting quantum Bose gases [41, 42]. Finally, it is fascinating to observe how different are the underlying mechanisms yielding to condensation transitions in haploid and diploid populations while both mechanisms have been selected by nature for their evolutionary advantages.

V. CONCLUSIONS

In conclusion we have studied a multiloci evolutionary dynamics in sexually reproducing diploid populations in which random genetic recombination tends to increase genetic variation while natural selection tends to reduce

it. The multiloci evolution is driven by a fitness function defined on an epistatic network of genetic loci. We have found that the stationary states of this evolutionary dynamics are multiple, and depend on the initial condition of the population. Unexpectedly, we have found that the joint distributions of allelic states at linked loci, can, at stationary state, be expressed in terms of a Bose-Einstein distribution with the "energy levels" depending on the network of epistatic interactions between genetic loci. The relation of the joint distributions with the Bose-Einstein distribution allows us to define a possible condensation phase transition in the universality class of the Bose-Einstein condensation. Below this condensation phase transition a finite fraction of pairs of genetic loci is fixed in the population and the number of polymorphic loci is strongly reduced. Therefore we propose here the Bose-Einstein condensation of genetic loci as a possible mechanism contributing to the reduction in genetic variation within a species.

In the future it is promising to include in this model the role of mutations (that increase genetic variation in the population), finite populations (that contribute to the existence of linkage disequilibrium) and the adaptive nature of the epistatic network. Moreover, we plan in future works to include in the model the possibility for a variable crossover rate and to go beyond the assumption of a locally tree-like epistatic network. Finally it would be interesting to characterize further the relation between the evolutionary dynamics and quantum mechanics [46, 47] by investigating the role of condensation transitions present in evolution [6–13] belonging to the Bose-Einstein universality class.

Appendix A: Calculations of marginals using the cavity method (Bethe-Peierls approximation)

In the hypothesis that the epistatic network is locally tree-like we look for solution to the evolutionary equation (5) of the form given by

$$P(\{x\}) = \prod_{\langle i,j \rangle} b_{ij}(x_i, x_j) \quad (\text{A1})$$

with $b_{ij}(x_i, x_j)$ to be determined by Eq. (5). The marginal frequency $p_{ij}(x_i, x_j)$ of a pair of linked loci is defined as

$$p_{ij}(x'_i, x'_j) = \sum_{\{x\}} P(\{x\}) \delta(x_i, x'_i) \delta(x_j, x'_j). \quad (\text{A2})$$

If we assume to know the functions $b_{ij}(x_i, x_j)$ the marginal frequencies (A2) can be calculated by the cavity method [37–40] (or the Bethe Peierls approximation) exactly valid on locally tree-like networks. For completeness we describe here the fundamentals of the cavity method that will be used in the following derivation of the stationary state solution of the evolutionary dynamics defined in (5). Let us work in the hypothesis that

the network is really a tree. Consider a known distribution function of the type (A1) with given functions $b_{ij}(x_i, x_j)$. Our purpose is to calculate in an efficient way the marginals $p_{ij}(x_i, x_j)$ defined in (13). These distributions can be calculated by a simple iterative procedure. Let us introduce the restricted partition functions $Z_{j|i}(x_j)$ of the sub-trees $\mathcal{T}_{j|i}$ rooted in the node $j \in \partial i$ and not including node i . These subtrees are marked by dashed lines in Fig. 6. The restricted partition functions $Z_{j|i}(x_j)$ is defined as

$$Z_{j|i}(x_j) = \sum_{\{x_\ell\}} \prod_{\ell, \ell' \in \mathcal{T}_{j|i}} b_{\ell, \ell'}(x_\ell, x_{\ell'}), \quad (\text{A3})$$

$\ell \in \mathcal{T}_{j|i} \setminus j$

where the sum is performed over all the variables x_ℓ associated with the nodes ℓ of the sub-tree $\mathcal{T}_{j|i}$ except for the variable x_j . Using this definition and the assumption that the network is locally a tree (see Fig. 7), it is easy to prove that the marginal distributions $p_{ij}(x_i, x_j)$ defined in (13) are given by

$$p_{ij}(x_i, x_j) = b_{ij}(x_i, x_j) Z_{j|i}(x_j) Z_{i|j}(x_i). \quad (\text{A4})$$

In order to calculate the restricted partition functions $Z_{j|i}(x_j)$ we use the following recursive equation, that expresses the relation between restricted partition functions of nested subtrees,

$$Z_{j|i}(x_j) = \prod_{k \in \partial j \setminus i} \sum_{x_k} b(x_j, x_k) Z_{k|j}(x_k). \quad (\text{A5})$$

These recursive equations are sufficient to define the full set of restricted partition functions $Z_{j|i}(x_j)$ within a constant that must be fixed by the normalization conditions

$$\sum_{x_i, x_j} p_{ij}(x_i, x_j) = 1. \quad (\text{A6})$$

The cavity method is proved to be exact not only on trees but also on locally tree-like networks. Nevertheless it also generally used for networks with short loops as long as the recursive equations (A5) have a solution. We can extend this formalism to generic distributions defined on locally tree-like networks in which each node is associated to more than one variable. Let us for example consider the case of the distribution function $\mathcal{P}(\{x^A, x^B\})$ defined in terms of the distribution $P(\{x\})$ and is given by

$$\mathcal{P}(\{x^A, x^B\}) = \left[\frac{W(\{x^A, x^B\}) P(\{x^A\}) P(\{x^B\})}{\langle W \rangle} \right]. \quad (\text{A7})$$

Also this distribution function, like the distribution $P(\{x\})$, is defined on a tree, but in this case to each node i , where two variables are associated: x_i^A and x_i^B . Assuming that the distribution $P(\{x\})$ is given by Eq. (A1), when the functions $b_{ij}(x_i, x_j)$ are known, we can

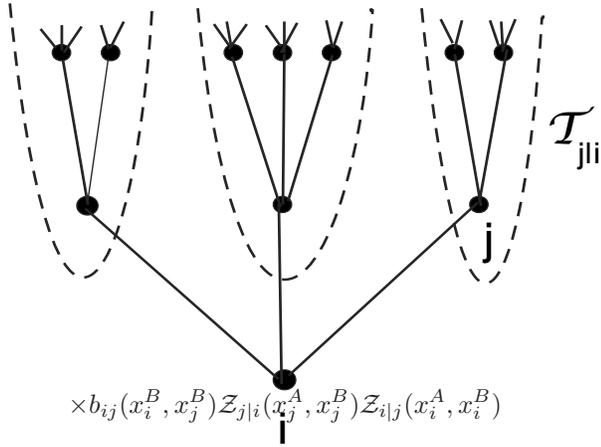


FIG. 6: Iteration tree for the calculation of the partition function. The sub-trees rooted in all $j \in \partial i$ are marked with dashed lines

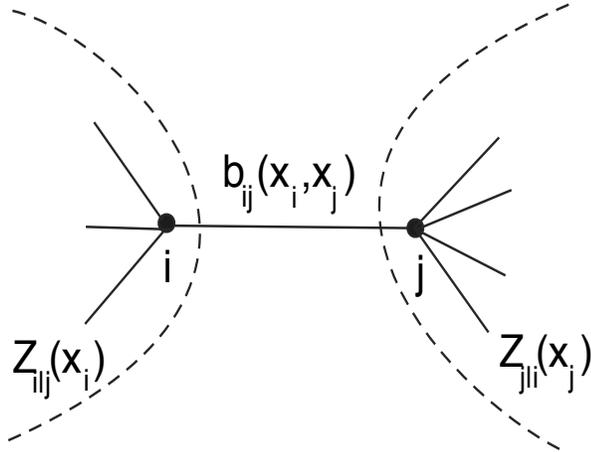


FIG. 7: The calculation of the marginal distribution $p_{ij}(x_i, x_j)$ can be expressed in terms on the restricted partition functions of the trees $\mathcal{T}_{i|j}$ and $\mathcal{T}_{j|i}$ according to Eq. (A4).

write the distribution $\mathcal{P}(\{x^A, x^B\})$ as a product of terms depending on indices (i, j) of linked pair of nodes according to the expression,

$$\mathcal{P}(\{x^A, x^B\}) = \prod_{\langle i, j \rangle} \frac{e^{-\beta U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)}}{\langle W \rangle} \times \prod_{\langle i, j \rangle} b_{ij}(x_i^A, x_j^A) b_{ij}(x_i^B, x_j^B). \quad (\text{A8})$$

Therefore, proceeding as in the previous case, we can use the cavity method and define the restricted partition functions $\mathcal{Z}_{j|i}(x_j^A, x_j^B)$ defined on the subtrees $\mathcal{T}_{j|i}$ and determined within a constant by the recursive equations

$$\mathcal{Z}_{j|i}(x_j^A, x_j^B) = \prod_{k \in \partial j \setminus i} \sum_{x_k^A, x_k^B} e^{-\beta U_{ik}(x_j^A, x_k^A, x_j^B, x_k^B)} \times b_{ik}(x_j^A, x_k^A) b_{ik}(x_j^B, x_k^B) \mathcal{Z}_{k|j}(x_k^A, x_k^B). \quad (\text{A9})$$

Finally the marginals $m_{ij}(x_i, x_j, x_i', x_j')$ of the probability distribution $\mathcal{P}(\{x^A, x^B\})$ are defined as

$$m_{ij}(x_i, x_j, x_i', x_j') = \sum_{\{x^A\}, \{x^B\}} \mathcal{P}(\{x^A, x^B\}) \delta(x_i^A, x_i) \times \delta(x_i^B, x_i') \delta(x_j^A, x_j) \delta(x_j^B, x_j') \quad (\text{A10})$$

and are given in terms of the restricted partition functions according to the following relation

$$m_{ij}(x_i^A, x_i^B, x_j^A, x_j^B) = \frac{1}{Z} e^{-\beta U_{ij}(x_i^A, x_j^A, x_i^B, x_j^B)} b_{ij}(x_i^A, x_j^A) \quad (\text{A11})$$

where the normalization constant Z can be calculated by imposing the normalization conditions

$$\sum_{x_i^A, x_j^A, x_i^B, x_j^B} m_{ij}(x_i^A, x_j^A, x_i^B, x_j^B) = 1. \quad (\text{A12})$$

Appendix B: Characterization of the steady state solution

The stationary states of Eq. (5) are given by the solutions to the equation

$$P(\{x\}) = M_{\{x\}} \left[\frac{W(\{x^A, x^B\}) P(\{x^A\}) P(\{x^B\})}{\langle W \rangle} \right]. \quad (\text{B1})$$

If the network of epistatic interactions is locally tree-like, we can find the exact solution of Eq. (B1) using a self-consistent argument [10] combined with the cavity method [37–40].

In our self-consistent assumption we suppose to know the functions $b_{ij}(x_i, x_j)$ determining the distribution $P(\{x\})$ given by Eq. (A1) and the distribution $\mathcal{P}(x^A, x^B)$ defined in Eqs. (A7) and (A8), both present in Eq. (B1). If we suppose to know the functions $b_{ij}(x_i, x_j)$ we can evaluate the marginal distributions $p_{ij}(x_i, x_j)$ and $m_{ij}(x_i, x_j, x_i', x_j')$ by the cavity method as described in the previous section. Finally imposing the stability condition (B1), on the marginal distributions, taking into account for the free recombination operator $M_{\{x\}}$ we get

$$b_{ij}(x_i, x_j) Z_{i|j}(x_i) Z_{j|i}(x_j) = \frac{1}{2} \sum_{x_i^B, x_j^B} m_{ij}(x_i, x_j, x_i^B, x_j^B) + \frac{1}{2} \sum_{x_i^B, x_j^A} m_{ij}(x_i, x_j^A, x_i^B, x_j) \quad (\text{B2})$$

The first term in the right hand side of Eq. (B2) comes from the probability that both allelic states i and j derive from a single parent. The second term in the right hand side of Eq. (B2), instead, describes the probability of a cross-over of genetic information, i.e. the event that the two allelic states (i, j) of the new gamete originate from

different parental gametes. In Eq. (B2) we have used the fact that the marginals $m_{ij}(x_i, x_j, x'_i, x'_j)$ are symmetric, i.e.

$$m_{ij}(x_i, x_j, x'_i, x'_j) = m_{ji}(x'_i, x'_j, x_i, x_j) \quad (\text{B3})$$

as a consequence of the assumed symmetries of the fitness function given by (4). Using Eq. (A11) to express the marginals, explicitly taking into account of the dependence of the right hand side of Eq. (B2) on $b_{ij}(x_i, x_j)$, we can write Eq. (B2) as in the following

$$Zb_{ij}(x_i, x_j)Z_{i|j}(x_i)Z_{j|i}(x_j) = b_{ij}(x_i, x_j)F_{ij}(x_i, x_j) + G_{ij}(x_i, x_j). \quad (\text{B4})$$

where the functions $F_{ij}(x_i, x_j)$, $G_{ij}(x_i, x_j)$ are defined as

$$\begin{aligned} F_{ij}(x_i, x_j) &= \frac{1}{2} \sum_{x_i^B, x_j^B} \frac{1}{b_{ij}(x_i, x_j)} m_{ij}(x_i, x_j, x_i^B, x_j^B) \\ &\quad + \frac{1}{2} \frac{1}{b_{ij}(x_i, x_j)} m_{ij}(x_i, x_j, x_i, x_j), \\ G_{ij}(x_i, x_j) &= \frac{1}{2} \sum_{x_j^A, x_i^B} m_{ij}(x_i, x_j^A, x_i^B, x_j) \\ &\quad \times [1 - \delta(x_i, x_i^B)\delta(x_j^A, x_j)]. \end{aligned} \quad (\text{B5})$$

These functions can be calculated by the cavity method in terms of the restricted partition functions $\mathcal{Z}_{i|j}(x_i)$ satisfying Eq.(A9) as expressed by the following equations,

$$\begin{aligned} F_{ij}(x_i, x_j) &= \frac{1}{2} \sum_{x_i^B, x_j^B} e^{-\beta U_{ij}(x_i, x_j, x_i^B, x_j^B)} b_{ij}(x_i^B, x_j^B) \\ &\quad \times \mathcal{Z}_{j|i}(x_j, x_j^B) \mathcal{Z}_{i|j}(x_i, x_i^B) + \\ &\quad + \frac{1}{2} e^{-\beta U_{ij}(x_i, x_j, x_i, x_j)} b_{ij}(x_i, x_j) \\ &\quad \times \mathcal{Z}_{j|i}(x_j, x_j) \mathcal{Z}_{i|j}(x_i, x_i) \\ G_{ij}(x_i, x_j) &= \frac{1}{2} \sum_{x_j^A, x_i^B} e^{-\beta U_{ij}(x_i, x_j^A, x_i^B, x_j)} b_{ij}(x_i, x_j^A) \\ &\quad \times b_{ij}(x_i^B, x_j) \mathcal{Z}_{j|i}(x_j^A, x_j) \mathcal{Z}_{i|j}(x_i, x_i^B) \times \\ &\quad \times [1 - \delta(x_i, x_i^B)\delta(x_j^A, x_j)], \\ Z &= \sum_{x_i, x_j} b_{ij}(x_i, x_j) F_{ij}(x_i, x_j) + G_{ij}(x_i, x_j). \end{aligned} \quad (\text{B6})$$

The Equations (B4), can be seen as a set of equations able to determine self-consistently the functions $b_{ij}(x_i, x_j)$ closing the self-consistent argument. The coupled equations (A5), (A6), (A9), (B4) and (B6) provide the solution for the stationary state of the multiloci evolution. These cavity equations will in general lead to multiple solutions corresponding to the multiplicity of possible steady states of the studied evolutionary dynamics.

Appendix C: Bose-Einstein distribution

We want here to comment on the structure of the stationary distribution found by the solution of the multiloci

evolution provided in the previous section. Solving Eqs. (B4) for $b_{ij}(x_i, x_j)$ yields

$$b_{ij}(x_i, x_j) = \frac{G_{ij}(x_i, x_j)}{ZZ_{i|j}(x_i)Z_{j|i}(x_j) - F_{ij}(x_i, x_j)}, \quad (\text{C1})$$

as long as $[ZZ_{i|j}(x_i)Z_{j|i}(x_j) - F_{ij}(x_i, x_j)] > 0$. Let us for the moment assume that this last condition is always satisfied and relate to Appendix D for the study of the solution when the mentioned condition is not met. We observe that the probability $p_{ij}(x_i, x_j)$ is given by

$$p_{ij}(x_i, x_j) = \frac{G_{ij}(x_i, x_j)Z_{i|j}(x_i)Z_{j|i}(x_j)}{ZZ_{i|j}(x_i)Z_{j|i}(x_j) - F_{ij}(x_i, x_j)}. \quad (\text{C2})$$

The stationary solution (C1) can be also written as

$$b_{ij}(x_i, x_j) = \frac{G_{ij}(x_i, x_j)/F_{ij}(x_i, x_j)}{e^{\beta[\epsilon_{ij}(x_i, x_j) - \mu]} - 1} \quad (\text{C3})$$

valid for $Ze^{\epsilon_{ij}(x_i, x_j)} - 1 > 0$, where $\epsilon_{ij}(x_i, x_j)$ and μ are defined as

$$\epsilon_{ij}(x_i, x_j) - \mu = \frac{1}{\beta} \ln \left[\frac{ZZ_{i|j}(x_i)Z_{j|i}(x_j)}{F_{ij}(x_i, x_j)} \right]. \quad (\text{C4})$$

Using the relation (A4) and the Eq.s (C1) – (C3), we derive the marginal probability $p_{ij}(x_i, x_j)$, of linked pairs of loci (i, j),

$$\begin{aligned} p_{ij}(x_i, x_j) &= b_{ij}(x_i, x_j)Z_{i|j}(x_i)Z_{j|i}(x_j) \\ &= \frac{1}{Z} G_{ij}(x_i, x_j) \{1 + n_B[\epsilon_{ij}(x_i, x_j)]\} \end{aligned} \quad (\text{C5})$$

with $n_B[\epsilon_{ij}(x_i, x_j)]$ indicating the Bose distribution [41] associated with "energy level" $\epsilon_{ij}(x_i, x_j)$, i.e.

$$n_B[\epsilon_{ij}(x_i, x_j)] = \frac{1}{e^{\beta[\epsilon_{ij}(x_i, x_j) - \mu]} - 1}. \quad (\text{C6})$$

Since the distribution of $P(\{x\})$ is normalized, μ must satisfy, for every pair of linked loci (i, j), the normalization condition

$$1 = \sum_{x_i, x_j} p_{ij}(x_i, x_j). \quad (\text{C7})$$

Using the expression (C5) for the marginal distribution $p_{ij}(x_i, x_j)$ we arrive at a set of equations,

$$1 = \frac{1}{Z} \sum_{x_i, x_j} G_{ij}(x_i, x_j) + \frac{1}{Z} \sum_{x_i, x_j} G_{ij}(x_i, x_j) n_B[\epsilon_{ij}(x_i, x_j)] \quad (\text{C8})$$

valid for every pair of linked loci (i, j). Summing Eq. (C8) over every pair of linked loci (i, j) we obtain

$$1 = \frac{1}{Z} \int d\epsilon g_\beta(\epsilon) [1 + n_B(\epsilon)] \quad (\text{C9})$$

where $n_B(\epsilon)$ is the Bose-Einstein distribution (C6) and $g_\beta(\epsilon)$ is given by

$$g_\beta(\epsilon) = \lim_{\Delta\epsilon \rightarrow 0} \frac{1}{\Delta\epsilon} \frac{1}{L} \sum_{i,j \in \partial i} \sum_{x_i, x_j} G_{i,j}(x_i, x_j) \times \chi_{\Delta\epsilon}(\epsilon_{ij}(x_i, x_j) - \epsilon) \quad (\text{C10})$$

with $\chi_\delta(x) = 1$ if $|x| \leq \delta$ and $\chi_\delta(x) = 0$ otherwise. Therefore the average fitness of the evolving population can be expressed in terms of an integral over a Bose-Einstein distribution with the "energy levels" to be determined self-consistently by the cavity method.

Appendix D: Condensation transition

An unexpected and new phenomenon can occur in this evolutionary process. Due to the fact that the joint probability of pairs of allelic states can be expressed in terms of a Bose-Einstein distribution we can predict that in this evolutionary dynamics a condensation in the same universality class as the Bose-Einstein condensation might occur. In a quantum Bose gas [41], a Bose-Einstein condensation is a phase transition at a critical value of the inverse temperature β_c such that for $\beta > \beta_c$ a finite fraction of the total number of particles is found in the ground state. The equivalent of this phase transition for the evolutionary dynamics described in this paper occurs when a finite fraction of pairs of loci gets fixed in given allelic configurations. Therefore, when this phenomenon occurs in an evolving diploid population, the number of polymorphic pairs of loci is reduced by a finite fraction.

Let us consider the case in which a pair of loci is fixed in the population, i.e.

$$p_{ij}(x_i, x_j) = \delta(x_i, x_i^*) \delta(x_j, x_j^*). \quad (\text{D1})$$

We want to prove that this condition is equivalent to the condition

$$\epsilon_{ij}(x_i^*, x_j^*) = \mu. \quad (\text{D2})$$

Given Eq. (16), we derive from the definition (B5) that the function $G_{ij}(x_i, x_j)$ is a constant and equal to zero, i.e. $G_{ij}(x_i, x_j) = 0, \forall (x_i, x_j)$. This result is evident if we observe that the only contributions to $G_{ij}(x_i, x_j)$, defined in (B5), are given by different pairs of allelic states (x_i, x_j) in the two parental gametes. Since we have assumed that all gametes have the same pair of allelic states in the genetic loci (i, j) , $G_{ij}(x_i, x_j) = 0 \forall (x_i, x_j)$. Inserting this result in Eq. (B4) we get the following relations

$$\begin{aligned} b_{ij}(x_i, x_j) &= 0 \text{ if } (x_i, x_j) \neq (x_i^*, x_j^*) \\ \epsilon_{ij}(x_i, x_j) &= \mu \text{ if } (x_i, x_j) = (x_i^*, x_j^*). \end{aligned} \quad (\text{D3})$$

Similarly it is easy to prove that Eq. (D2) implies Eq. (D1). Therefore, if we want to describe fixed genetic loci, we have to modify Eq. (C5), for the marginal probability $p_{ij}(x_i, x_j)$ according to the following expression,

$$p_{ij}(x_i, x_j) = \begin{cases} 1 & \text{if } \epsilon_{ij}(x_i, x_j) = \mu \\ \frac{1}{Z} G_{ij}(x_i, x_j) \{1 + n_B[\epsilon_{ij}(x_i, x_j)]\} & \text{otherwise} \end{cases}$$

Accordingly, expressions (C1) and (C3) for $b_{ij}(x_i, x_j)$ have to be modified in order to take into account the possibility that a pair of loci gets fixed. Therefore we have

$$b_{ij}(x_i, x_j) = \begin{cases} \frac{G_{ij}(x_i, x_j)}{F_{ij}(x_i, x_j)} n_B[\epsilon_{ij}(x_i, x_j)] & \text{if } \epsilon_{ij}(x_i, x_j) > \mu \\ [Z_{ij}(x_i) Z_{j|i}(x_j)]^{-1} & \text{if } \epsilon_{ij}(x_i, x_j) = \mu. \end{cases}$$

The set of equations (C8) consistent with the normalization condition (C7) is therefore modified and takes the form

$$\begin{aligned} (1 - \nu_{ij}) &= \frac{1}{Z} \sum_{x_i, x_j} G_{ij}(x_i, x_j) + \\ &+ \frac{1}{Z} \sum_{\substack{x_i, x_j \\ \epsilon_{ij}(x_i, x_j) > \mu}} G_{ij}(x_i, x_j) n_B[\epsilon_{ij}(x_i, x_j)] \end{aligned} \quad (\text{D4})$$

valid for every pair of linked loci (i, j) . In Eq. (D4) the matrix elements ν_{ij} are taken such that $\nu_{ij} = 1$ if a pair of configurations (x_i^*, x_j^*) exists such that $\epsilon_{ij}(x_i^*, x_j^*) = \mu$, otherwise we have $\nu_{ij} = 0$. The study of the normalization equation (D4) will define if and when the number of pairs of fixed loci becomes extensive. In the presence of a negligible fraction of fixed pairs of loci, averaging (D4) over all pairs of links we get the equation (C9). For the values of the evolutionary pressure for which Eq. (C9) cannot be satisfied, a finite fraction ν_0 of genetic loci is fixed and the conservation equation (C9) has to be modified according to

$$(1 - \nu_0) = \frac{1}{Z} \int_{\epsilon > \mu} d\epsilon g_\beta(\epsilon) \left[1 + \frac{1}{e^{\beta(\epsilon - \mu)} - 1} \right] \quad (\text{D5})$$

with $g_\beta(\epsilon)$ given by (C10) and ν_0 defined as

$$\nu_0 = \frac{1}{2L} \sum_{i,j \in \partial i} \sum_{x_i, x_j} \chi_0[\epsilon_{ij}(x_i, x_j) - \mu], \quad (\text{D6})$$

where $\chi_\delta(x) = 1$ if $|x| \leq \delta$ and $\chi_\delta(x) = 0$ otherwise. As a function of the evolutionary pressure, a condensation transition can occur between a "non condensed phase" in which all the genetic loci are polymorphic, and a "condensed phase" in which only a fraction of the genetic loci is polymorphic. This phase transition is in the universality class of the Bose-Einstein condensation and it can be compared with other condensation phase transitions in haploid and diploid evolution [6, 7, 9, 11, 13].

Acknowledgments

The authors thank Paola Ricciardi-Castagnoli for interesting comments and discussions.

-
- [1] R. A. Fisher, *The Genetical Theory of Natural Selection*, (Clarendon Press, Oxford, 1930).
- [2] M. Kimura, *J. of Appl. Probab.* **1**, 177 (1964).
- [3] G. Sella and A. E. Hirsh, *Proc. Natl. Acad. Sci. USA* **102**, 9541 (2005).
- [4] U. Gerland, J. D. Moroz and T. Hwa, *Proc. Natl. Acad. Sci. USA* **99**, 12015 (2002).
- [5] J. Monod, *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, (William Collins Sons & Co., Glasgow, 1972).
- [6] M. Eigen, *Naturwiss.* **58**, 465 (1971).
- [7] S. Franz, L. Peliti, *J. Phys. A* **30**, 4481 (1997).
- [8] M. A. Nowak, *Evolutionary Dynamics*, (Harvard University Press, Cambridge, MA, 2006).
- [9] J. F. C. Kingman, *J. Appl. Probab.* **15**, 1 (1978).
- [10] G. Bianconi and A. L. Barabási, *Phys. Rev. Lett.* **86**, 5632 (2001).
- [11] S. N. Coppersmith, R. D. Blanck and L. P. Kadanoff, *Jour. Stat. Phys.* **97**, 1999 (2004).
- [12] G. Bianconi, L. Ferretti and S. Franz, *EPL* **87**, 28001 (2009).
- [13] R. A. Neher and B. I. Shraiman, *Proc. Natl. Acad. Sci. USA* **106**, 6866 (2009).
- [14] Eric S. Lander et al. *Nature* **409**, 860 (2001).
- [15] J. C. Venter et al., *Science* **291**, 1304 (2001).
- [16] A.-L. Barabási and Z. Oltvai, *Nat. Rev. Genet.* **5**, 101 (2004).
- [17] E. Ben-Naim, H. Frauenfelder and A. Toroczkai, *Complex Networks Lecture Notes in Physics* 650, (Springer-Verlag, 2004).
- [18] K. Sneppen and G. Zocchi, *Physics in molecular biology*, (Cambridge University Press, Cambridge, England, 2005).
- [19] S. Bornholdt, *Science* **310**, 449 (2005).
- [20] R. Albert, *Jour. of Cell Sci.* **118**, 4947 (2005).
- [21] U. Alon, *An introduction to system biology: design principles of biological circuits*, (Chapman & Hall, London, 2007).
- [22] M. Slatkin, *Nature Rev. Genet.* **9**, 477 (2008).
- [23] D. Segré, A. Deluna, G. M. Church and R. Kishony, *Nature Gen.* **37**, 77 (2009).
- [24] M. Costanzo M et al., *Science* **327**, 425 (2010).
- [25] E. Ravasz, A.L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabási, *Science* **297**, 1551 (2002).
- [26] H. Jeong, S. P. Mason, A. L. Barabási and Z. N. Oltvai, *Nature (london)* **411**, 41 (2001).
- [27] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002).
- [28] A. Vazquez, A. Flammini, A. Maritan and A. Vespignani, *Nature Biotech.* **21**, 697 (2003).
- [29] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási *Nature (London)* **407**, 651 (2000).
- [30] C. I. Jones et al., *Blood* **114**, 1406 (2009).
- [31] D. E. Reich et al., *Nature (London)* **411**, 199 (2001).
- [32] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, (Cambridge University Press, Cambridge, England, 1998).
- [33] J. H. Gillespie, *Population Genetics: A concise Guide*, (John Hopkins University Press, Baltimore, MD, 2004).
- [34] D. Hartl and A. G. Clark, *Principles of population genetics*, (Sinauer Associates Inc. Publisher, Sunderland, 2007).
- [35] K. Jain and J. Krug, in *Structural Approaches to Sequence Evolution*, eds. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo (Springer-Verlag, Berlin, 2007).
- [36] C. Darwin, *On the Origin of Species by means of natural selection*, (Oxford University Press, New York, 1859).
- [37] M. Mézard and G. Parisi, *Eur. Phys. Jour.* **20**, 217 (2001).
- [38] J. S. Yedidia, W. T. Freeman and Y. Weiss, in: *Exploring artificial Intelligence in the New Millennium* (Science and Technology Books, 2003).
- [39] A. Hartmann and M. Weigt, *Phase Transitions in Combinatorial Optimization Problems*, (Wiley-VCH, Weinheim, 2005).
- [40] M. Mézard and A. Montanari, *Information, Physics and Computation*, (Oxford University Press, Oxford, 2009).
- [41] K. Huang *Statistical Mechanics*, (John Wiley and Sons, 1987).
- [42] C. J. Pethick and H. Smith *Bose-Einstein Condensation in Diluted Gases*, (Cambridge University Press, Cambridge, 2001).
- [43] J. Maynard Smith and J. Haigh, *Genetics Research* **23**, 23 (1974).
- [44] L. Peliti and U. Bastolla, *CR Acad. Sci. III* **317**, 371 (1994).
- [45] P. A. P. Moran, *Ann. Hum. Genet.* **27**, 383 (1964).
- [46] P. C. W. Davies, *BioSystems* **78**, 69 (2004).
- [47] S. Lloyd, *Nature Physics* **5**, 164 (2009).