



## OPEN

SUBJECT AREAS:  
BIOLOGICAL PHYSICS  
POPULATION GENETICS  
COMPUTATIONAL MODELS  
STATISTICAL PHYSICS,  
THERMODYNAMICS AND  
NONLINEAR DYNAMICS

# Tipping points and early warning signals in the genomic composition of populations induced by environmental changes

Jacobo Aguirre<sup>1,2,3</sup> & Susanna Manrubia<sup>1,2,3</sup>

<sup>1</sup>Centro de Astrobiología (INTA-CSIC), ctra. de Ajalvir km 4, 28850 Torrejón de Ardoz, Madrid, Spain, <sup>2</sup>Centro Nacional de Biotecnología, CSIC, c/Darwin 3, 28049 Madrid, Spain, <sup>3</sup>Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain.

Received  
2 January 2015

Accepted  
12 March 2015

Published  
12 May 2015

Correspondence and requests for materials should be addressed to S.M. (smanrubia@cnb.csic.es)

We live in an ever changing biosphere that faces continuous and often stressing environmental challenges. From this perspective, much effort is currently devoted to understanding how natural populations succeed or fail in adapting to evolving conditions. In a different context, many complex dynamical systems experience critical transitions where their dynamical behaviour or internal structure changes suddenly. Here we connect both approaches and show that in rough and correlated fitness landscapes, population dynamics shows flickering under small stochastic environmental changes, alerting of the existence of tipping points. Our analytical and numerical results demonstrate that transitions at the genomic level preceded by early-warning signals are a generic phenomenon in constant and slowly driven landscapes affected by even slight stochasticity. As these genomic shifts are approached, the time to reach mutation-selection equilibrium dramatically increases, leading to the appearance of hysteresis in the composition of the population. Eventually, environmental changes significantly faster than the typical adaptation time may result in population extinction. Our work points out several indicators that are at reach with current technologies to anticipate these sudden and largely unavoidable transitions.

The identification of indicators signalling the proximity of a critical transition in the dynamics of natural systems is an emerging field of research<sup>1</sup>. Early-warning signals have been described in widely different systems which, however, share important architectural features. These include a global network of connections between components and a non-linear response to change promoted by positive feedbacks<sup>2</sup>. Often, those studies rely on simple models that have educated our intuition and clearly delineated the framework where tipping points occur, many of them focusing on climate dynamics<sup>3,4</sup>. The conceptual tools at hand have been applied to characterize *a posteriori* sudden shifts of state in ecosystems<sup>5</sup>, physiology, and finance<sup>2</sup>, and are beginning to be used to identify tipping points in *in vitro* natural populations<sup>6</sup>. In all those examples, the system dynamics proceeds under the effect of external changes that implicitly modify the fitness landscape. Indeed, the fitness value of a given population is context dependent, and thus varies concomitantly with the environment: eventually, sudden changes of state occur. Among several other well-documented examples in ecology we find the loss of coverage by charophyte vegetation experienced by lake Veluwe in response to the increase in phosphorus concentration<sup>7</sup>. The sudden decrease observed at a critical concentration in the early 1970s was later reverted in the 1990s as the concentration of phosphorus decreased, though that occurred through a hysteresis cycle. In single populations of unicellular organisms<sup>6</sup>, an instability separating a parameter region leading to a stable fixed point of high population density from a region ushering in extinction appears as a consequence of variations in population density. That experiment can be understood as a case of a slowly driven landscape that causes systematic decreases in the population density. In a fitness landscape representation of this second example, the lower the population density the lower its fitness. The population is pushed towards a valley and recovers with increasing difficulty until, at a critically low density, the attractor of the dynamics shifts to the extinct state.

Recent studies posit that smooth alterations of the environment might be responsible for some of the drastic state shifts that have been reported in our planet's history, and warn about the repetition of this phenomenon in the nearby future<sup>2,8,9</sup>. Sudden shifts in the states of evolving populations interrupting long periods of stagnancy have been described at the level of species<sup>10</sup>, bacteria<sup>11</sup> and viral genomes<sup>12</sup>. Several models have interpreted those



transitions as jumps along evolutionary optimization, with stasis periods due to entropic or fitness barriers that the population has difficulty in overcoming<sup>13–15</sup>, or as selective sweeps leading to a new phenotype<sup>12</sup>. Here we analyse these transitions in realistic, slowly changing fitness landscapes of genomes, and show that it is a generic phenomenon. Though most populations experience sudden genomic shifts in out-of-equilibrium situations, as those described above, this behavior also occurs when populations are allowed to attain mutation-selection equilibrium between perturbations (either stochastic or driven), as we here show. Several indicators quantify the nearness of the transition.

In Methods we present a general model to study population dynamics on stochastic landscapes. In particular, we devise a representative example of a broad class of models for evolving populations in fixed environments. The fitness landscape is generated through the NK model of Kauffman<sup>16</sup>, which yields landscapes of tunable ruggedness with properties characteristic of natural systems: multiple peaks, epistatic interactions, and local optima<sup>17</sup>. The landscape is a network of genomes mutually accessible through mutations. It is endowed with a fraction of lethal or strongly deleterious mutations, in agreement with observations<sup>18</sup>. Finally, we pay special attention to the mathematical description of the population dynamics over the fitness landscape, and the relevant quantities that characterize it.

Evidence in favour of a direct effect of environmental conditions in fitness landscapes is mounting<sup>19–21</sup>. In Results we analyse time-varying fitness landscapes by implementing environmental perturbations through the addition of stochastic noise to the fitness value of genomes. We focus on two types of varying environments: (i) a constant fitness landscape affected by stochastic fluctuations; and (ii) a driven fitness landscape. The latter is motivated by the fact that driving forces in the history of our planet leading to global changes, such as temperature, concentration of oxygen, carbon dioxide or hydrogen sulphide, changed smoothly but monotonically — in many cases almost linearly — with time before tipping points were reached<sup>2,8</sup>. We prove analytically and characterize numerically that sharp transitions at the genomic level preceded by early-warning signals are a generic phenomenon in constant and slowly driven landscapes affected by even slight stochasticity, and focus on the robustness of the phenomenon described. In the Discussion we analyse the generality of the results obtained, their influence in our understanding of complex ecological systems, and present a protocol based on current high-throughput sequencing technologies to anticipate these sudden genomic transitions whose consequences are not fully understood yet.

## Methods: General model for population dynamics on stochastic landscapes

**The stochastic fitness landscape.** A genome is defined as a sequence of length  $N$ . The state of each locus is taken from an alphabet of  $A$  letters. The space of genomes is a regular network of size  $m = A^N$  and dimension  $N$  where two nodes (or genomes) are connected by an undirected link if they differ in the state of only one locus. Each node has  $(A - 1)N$  neighbours or connections to other genomes. This quantity defines its degree. The maximum distance between two sequences is  $N$ . When  $A = 2$  the space of genomes is a hypercube of dimension  $N$ .

The NK model is based on the spin glass model of statistical physics<sup>22</sup>, and represents one of the most versatile models to map genotypes to rugged fitness landscapes<sup>16,23,24</sup>. It represents genomes with  $N$  loci (genes, nucleotides, aminoacids, etc.), that take one out of  $A$  different states (number of alleles, nucleotides, aminoacids, etc.) Epistasis, or interaction between loci, is defined here as an interaction of each locus with  $K$  other loci of the same genome. The larger  $K$ , the more rugged the landscape (see below).

The fitness value  $f_s$  of a genome  $s$  in a given fitness landscape  $\mathcal{L}$  is calculated as follows (for the sake of clarity we have presented a detailed example in section S1 of the Supplementary Information). First, we construct a matrix  $L$  of size  $A^{K+1} \times N$  whose elements  $L_{ij}$  are random numbers uniformly chosen between 0 and 1. Elements in column  $j$ ,  $L_{i=1\dots A^{K+1},j}$ , represent all possible contributions of locus  $j$  to the total fitness of the genome  $s$ . The actual contribution of locus  $j$ ,  $L_{i,j}$ , depends on its state and on those of the  $K$  other loci with which it interacts. For this reason, there are  $A^{K+1}$

values in each column, representing the  $A^{K+1}$  possible combinations of  $K+1$  elements taking  $A$  different states (and index  $i$  stands for the correct one for locus  $j$ ).

The fitness of sequence  $s$  is the average value of the fitness contributions of each of its loci,

$$f_s = \frac{\sum_{j=1}^N L_{i,j}}{N}. \quad (1)$$

Note that, due to the normalisation above, all fitness values satisfy  $0 \leq f_s \leq 1$ . As matrix  $L$  is stochastically generated, the corresponding fitness landscape  $\mathcal{L}$  is different every time we run the algorithm. Varying parameter  $K$  yields fitness landscapes with features akin to the properties measured in different natural genome spaces (DNA, RNA, proteins, etc.)<sup>25</sup>. Low values of  $K$  give rise to smooth Fujiyama landscapes, that is, landscapes with only one high peak and where similar genotypes show similar fitness values. High values of  $K$  describe rugged landscapes with many local maxima and where close genotypes may show very different fitness, the more different the higher  $K$ . In the upper limit  $K = N - 1$ , the fitness of each genotype is not correlated at all with that of any other genotype. Parameter  $K$  thus characterises landscape ruggedness.

Finally, the existence of sequences carrying lethal mutations, that is with fitness zero, cannot be discarded in realistic fitness landscapes: Natural landscapes are holey<sup>26</sup> and a significantly large fraction of all possible mutations leads to non-viable individuals<sup>18</sup>. These two features result in actual genome networks that are topologically heterogeneous.

Lethal mutations are introduced in the model as follows. First, we generate the NK-fitness  $(f_0)_s$  of each sequence  $s$  as explained above. Then, we introduce the lethality coefficient, a new parameter  $0 \leq f_l \leq 1$  such that those genotypes with  $(f_0)_s < f_l$  are defined as lethal and assigned fitness zero ( $f_s = 0$ ), and the remaining values are rescaled: if  $(f_0)_s > f_l$ , then  $f_s = (f_0)_s - f_l$ . A fraction of lethal genomes is so generated, whose number has to be calculated after the truncation of the landscape has been applied.

This procedure immediately yields a heterogeneous network, now formed by nodes with degree between 0 and the maximum value  $(A - 1)N$ . In the limit  $K = N - 1$ , the fitness of each sequence is independent of its neighbours and the lethal mutations so introduced would thus be randomly distributed. For any  $K < N - 1$  correlations exist and our procedure yields connected groups of genomes larger than randomly expected and also clustering in the distribution of lethal mutants. Both facts are in agreement with the existence of mutational hot spots (regions in the genome that tolerate mutations) and highly conserved sequences (regions where mutations likely cause deleterious effects and are eliminated by selection).

**Mathematical description of population evolutionary dynamics in a fixed fitness landscape.** The evolutionary dynamics of a population replicating and mutating in a networked, fixed fitness landscape, can be cast in mathematical form<sup>27,28</sup>. For the sake of completeness, here we present its main properties applied to the present situation.

A wide variety of evolutionary processes occurring in a network of  $m$  nodes can be mathematically described as

$$\vec{n}(g+1) = \mathbf{M}\vec{n}(g), \quad (2)$$

where  $\vec{n}(g)$  is a vector whose components are the population of individuals at each node at generation (or time)  $g$  and  $\mathbf{M}$ , with  $M_{ij} \geq 0$ , is a transition matrix of dimension  $m \times m$  that contains all details on the evolutionary process (note that  $m = A^N$  in our model).

$\mathbf{M}$  is a primitive matrix. For this reason, its largest eigenvalue is positive, it verifies that  $\lambda_1 > |\lambda_i|$ ,  $\forall i > 1$ , and the elements of its associated eigenvector  $\vec{u}_1$  are also positive. Therefore, the dynamics can be written as

$$\vec{n}(g) = \mathbf{M}^g \vec{n}(0) = \sum_{i=1}^m (\vec{n}(0) \cdot \vec{u}_i) \lambda_i^g \vec{u}_i, \quad (3)$$

where  $\vec{n}(0)$  is the initial condition and  $\vec{u}_i$  the  $i$ -th eigenvector of  $\mathbf{M}$ ,  $i = 1, \dots, m$ . Eq. (3) reveals that the system evolves towards an asymptotic state independent of the initial condition and proportional to the first eigenvector  $\vec{u}_1$ . Its associated eigenvalue  $\lambda_1$  yields the growth rate at the asymptotic equilibrium. If  $\vec{n}(g)$  is normalised such that  $\|\vec{n}(g)\| = 1$  after each iteration (the population is kept constant),  $\vec{n}(g) \rightarrow \vec{u}_1$  when  $g \rightarrow \infty$ .

As explained above, strictly speaking the number of generations elapsed before equilibrium is reached is infinity. However, there is a way we can define an effective time to equilibrium  $g_e$ . We define  $g_e$  as the number of generations required for

$\Delta(g_e) = \left\| \frac{\vec{n}(g_e)}{\|\vec{n}(g_e)\|} - \vec{u}_1 \right\| \leq \zeta$ , where  $\zeta \ll 1$  is a tolerance. In all cases where  $0 \leq \lambda_3 < \lambda_2$  and  $\vec{n}(0) \cdot \vec{u}_2 \neq 0$ ,  $g_e$  can be calculated from Eq. (3) and  $\Delta(g_e)$  through the excellent approximation

$$g_e \sim \frac{\ln \left| \frac{\vec{n}(0) \cdot \vec{u}_2}{\vec{n}(0) \cdot \vec{u}_1} \right| - \ln \zeta}{\ln \frac{\lambda_1}{\lambda_2}} \quad (4)$$

thanks to the exponentially fast suppression of the contributions of higher-order terms ( $\lambda_i \geq \lambda_{i+1}$ ,  $\forall i$ ; see<sup>27</sup>).



Let us focus on the description of the current model. The explicit form of the transition matrix  $M$  depends on the particular system studied, but it is always a function of the mutual accessibility of genotypes and their fitness. These two quantities are represented through (i) the adjacency matrix  $G$ , which encodes the regular topology of the space of genomes; elements of  $G$  are  $G_{ij} = 1$  if nodes  $i$  and  $j$  are connected and  $G_{ij} = 0$  otherwise, and (ii) diagonal matrix  $F$  containing the fitness value of each node obtained following the former section,  $F_{ij} = f_i \delta_{ij}$ , where  $i, j = 1, \dots, A^N$ .

For instance, a population whose individuals produce  $r$  offspring per generation that mutate with a probability  $\mu$  per genome and replication cycle, and which substitute the parental population, evolves according to the transition matrix

$$M = r(1 - \mu)FI + \frac{r\mu}{S}FG, \quad (5)$$

where  $S \equiv (A - 1)N$  is the total number of neighbours of each genome. Note that the effective reproduction rate of genome  $i$  is  $rf_i$  due to the fitness factor. However, it is important to emphasise that the only quantity capturing the growth rate of the population at equilibrium (and thus the replication rate of all sequences in the population) is  $\lambda_1$ , the largest eigenvalue of matrix  $M$ .

In the particular case studied here, we have chosen  $\mu = 1$  to highlight the role played by the second term, which contains all information on the interplay between fitness and topology. Also, mutation rates of the order of one change per genome and replication cycle are not rare<sup>29</sup>, so this situation is representative at least of fast mutating organisms.

In general, the growth of the population at equilibrium is given by the largest eigenvalue  $\lambda_1$ . Hence,  $\lambda_1 > 1$  implies exponential growth,  $\lambda_1 = 1$  stands for a constant population size, and  $\lambda_1 < 1$  means a shrinking population. Note that the  $NK$  model, by definition, produces values of  $F_{ij} < 1$ . Therefore,  $r$  should be large enough so as to avoid extinction. In general, indeed, fitness values are arbitrary as long as they are larger or equal than zero, so they admit a rescaling by any constant factor which, if large enough, would eventually yield  $\lambda_1 > 1$ . As explained above, this rescaling (here represented through  $r$ ) does neither modify the eigenvectors of matrix  $M$ , nor the dynamics of the population. In order to avoid maintaining a meaningless factor in front of our dynamical equations, and without loss of generality, we will study the case

$$M = FG. \quad (6)$$

In the case of the  $NK$  model, as discussed,  $\lambda_1 < 1$ , but we could simply rescale the height of the fitness landscape by a constant arbitrary factor (subsumed in  $r$ ) without causing any other quantitative change. For completeness, in section S2 of the Supplementary Information we will show that all the phenomenology obtained for Eq. (6) is even quantitatively enhanced for values of the mutation rate  $\mu < 1$  and using the more general Eq. (5).

The transition matrix defined by Eq. (6) is symmetrizable and all its eigenvalues  $\lambda_i$  are real<sup>27</sup>. However,  $M$  is not symmetric, because of its dependence with  $F$ . As a result, there is no simple relationship between the eigenvalues and eigenvectors of the symmetric matrix  $G$  and those of  $M$ : topology and fitness cannot be decoupled.

**Definition of relevant quantities in fixed fitness landscapes.** In addition to the quantities already introduced, as the number of generations to equilibrium  $g_e$ , and all those characterising the global dynamics as  $\lambda_i$  and  $\bar{u}_i$ , we introduce in the following a number of variables that quantify the evolutionary process and some important changes in its nature. Let us remark that though the quantities presented below have been defined at mutation-selection equilibrium, that is, for  $g \rightarrow \infty$ , they can be generalised for any time  $g$  by substituting  $\bar{u}_i$  by the corresponding state of the population  $\bar{u}(g)$ .

**Matrix of minimum distances between genomes.** The fitness landscape we are investigating has a fraction of sequences with zero fitness. They cause holes in the landscape, which correspond to regions that cannot be traversed by the population. We now define matrix  $D$  as a matrix of distances between genomes. Its elements  $D_{ij}$  yield the minimum path between sequences  $i$  and  $j$  when linked through the adjacency matrix but avoiding sequences with zero fitness. In other words, the value of  $D_{ij}$  is the minimum number of mutations that permit to move from genome  $i$  to genome  $j$  by single mutations through a path of positive fitness.

**Genetic distance at equilibrium.** In order to quantify the position of the population in the fitness landscape at equilibrium, we have introduced the genetic distance  $d_0$ . This distance evaluates how far, according to  $D$ , each of the genomes currently represented in the population is from an arbitrary reference point. As reference we have chosen sequence  $s^0 = \bar{0} = (0, 0, 0, \dots, 0)$ . The total distance is the sum over all individuals. In mathematical form,

$$d_0 = \frac{\sum_{i=1}^{A^N} (u_1)_i D_{i1}}{\sum_{i=1}^{A^N} (u_1)_i}. \quad (7)$$

The first eigenvector of the transition matrix  $M$ ,  $\bar{u}_1$ , represents the population at equilibrium and assigns the corresponding weight to each sequence (subindex  $i$  stands for the  $i$ -th component of the vector); matrix  $D$  has been ordered in such a way

that its first column contains the distances of all nodes in the landscape to sequence  $s^0$ . The normalised product of those two quantities defines the genetic distance  $d_0$ .

**Average fitness of the population at equilibrium.** The fitness value of node  $i$  in the landscape is  $f_i$ . Consequently, the fitness associated to such landscape is  $\vec{f} = (f_1, \dots, f_{A^N})$ , and the average fitness of the population at equilibrium,  $\bar{F}$ , is

$$\bar{F} = \frac{\vec{f} \cdot \bar{u}_1}{\sum_{i=1}^{A^N} (u_1)_i}. \quad (8)$$

**Definition of relevant quantities in time-varying fitness landscapes.** Up to here, we have focused on the dynamics of populations on fixed landscapes. However, in natural environments landscapes evolve in time. The following quantities measure the evolution of the population distribution over the space of genomes and that of the fitness landscape in a varying environment.

**Difference to initial equilibrium.** In a varying environment we might be interested in the difference between the state of the population at the equilibrium or the fitness landscape at a given environment  $\tau$  and in the original situation  $\tau = 1$ . The difference between the mutation-selection equilibrium of the population in the initial landscape and the equilibrium attained in a later environment  $\tau$  is quantified as

$$\Delta \bar{u}_1(\tau, 1) = 1 - \frac{\bar{u}_1(\tau) \cdot \bar{u}_1(1)}{\|\bar{u}_1(\tau)\| \|\bar{u}_1(1)\|}. \quad (9)$$

The difference between the initial fitness landscape and that at environment  $\tau$  is similarly calculated,

$$\Delta \vec{f}(\tau, 1) = 1 - \frac{\vec{f}(\tau) \cdot \vec{f}(1)}{\|\vec{f}(\tau)\| \|\vec{f}(1)\|}. \quad (10)$$

**Difference between subsequent equilibria.** Similarly to the former definitions, in a varying environment the difference of the state of the population at the equilibrium at two contiguous environments  $\tau$  and  $\tau + 1$  is precisely defined as

$$\Delta \bar{u}_1(\tau, \tau + 1) = 1 - \frac{\bar{u}_1(\tau) \cdot \bar{u}_1(\tau + 1)}{\|\bar{u}_1(\tau)\| \|\bar{u}_1(\tau + 1)\|}, \quad (11)$$

and the corresponding relative change in fitness landscape is

$$\Delta \vec{f}(\tau, \tau + 1) = 1 - \frac{\vec{f}(\tau) \cdot \vec{f}(\tau + 1)}{\|\vec{f}(\tau)\| \|\vec{f}(\tau + 1)\|}. \quad (12)$$

Note that the difference to initial condition (Eqs. (9) and (10)) and the difference between subsequent equilibria (Eqs. (11) and (12)) are defined in<sup>9,1</sup>, where zero and one stand respectively for a total similarity and an absolute dissimilarity between the composition of the population and the fitness landscape at the two compared environments.

## Results

Fitness values are strongly influenced by the environment. Fitness is a context-dependent quantity, and multiple causes modify the environmental conditions of ecosystems. Inspired by two common scenarios, we have implemented the following situations where fitness changes are effectively affected by the environment: (i) a stochastically varying fitness landscape and (ii) a driven fitness landscape.

**Stochastically varying fitness landscape.** We first consider a randomly fluctuating landscape. For this purpose, we calculate a starting fitness landscape  $\mathcal{L}_{\tau_0}$ , whose associated random matrix is  $L_{\tau_0}$ , in the way shown in Methods. The elements of the random matrices  $L_{\tau}$  associated to subsequent fitness landscapes  $\mathcal{L}_{\tau}$  are obtained as follows

$$(L_{\tau+1})_{ij} = (L_{\tau})_{ij} + \epsilon \eta_{ij}, \quad (13)$$

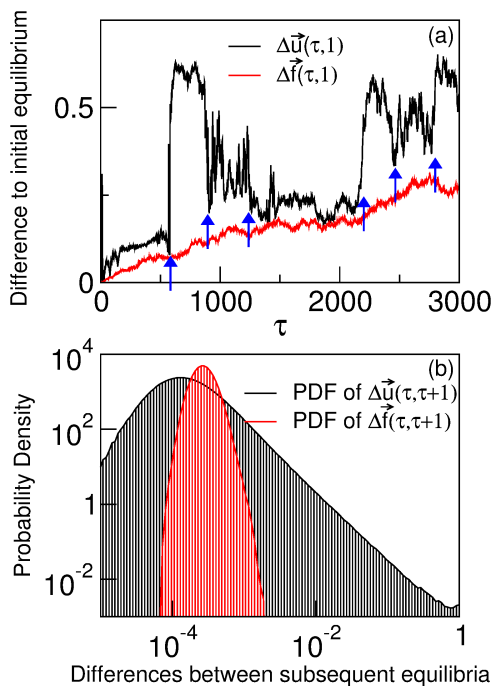
where  $\eta_{ij}$  is a random number uniformly distributed in  $[-1, 1]$  and  $\epsilon$  is the environmental noise intensity. We use the variable  $\tau$  to order the fitness landscapes where the population will evolve. Note that it is not a proper time, but a label used for convenience to indicate the succession of environments.



The fitness of all sequences is obtained as described in Methods. Our procedure maintains the expected correlation, given  $K$ , between the fitness of neighbouring genotypes, since the stochastic variation is added to the elements of matrix  $L_\tau$  and not to the values  $f_s$ .

In Fig. 1 we calculate the mutation-selection equilibrium ( $g \rightarrow \infty$ ) of a genomic population with sequences of length  $N = 8$  and a two-letter alphabet, for a fitness landscape  $\mathcal{L}_\tau$  that evolves stochastically as explained above. It turns out that the population distribution in subsequent environments can be deeply dissimilar to that of the initial environment for fitness landscapes affected by little change. Figure 1(a) plots the difference of the population distribution at environment  $\tau$  to the initial equilibrium,  $\Delta\vec{u}_1(\tau, 1)$  (Eq. (9)), and that of the fitness landscape  $\Delta\vec{f}(\tau, 1)$  (Eq. (10)). The environmental noise of small amplitude causes flickering in the genomic composition of populations even when the time scales of adaptation and environmental change are decoupled. This strongly non-linear response of the population is captured by the histograms of  $\Delta\vec{u}_1(\tau, \tau + 1)$  (Eq. (11)) and  $\Delta\vec{f}(\tau, \tau + 1)$  (Eq. (12)), the differences between the population and fitness state in the equilibria at contiguous environments  $\tau$  and  $\tau + 1$  (Fig. 1(b)). While the distribution of changes in the landscape is narrow, that of changes in the population state shows a power-law-like distribution, highlighting the generality of the abrupt transitions marked by arrows in Fig. 1(a).

But how general is the presence of these genomic shifts when we vary the parameters, and which is the reason for their appearance? In



**Figure 1 | Mutation-selection equilibria in a stochastically varying environment.** (a) Difference of the state of the population ( $\Delta\vec{u}_1(\tau, 1)$ , black) and the fitness landscape  $\Delta\vec{f}(\tau, 1)$ , (red) to the initial environment  $\tau_0 = 1$ . The fitness landscape drifts randomly as small stochastic perturbations accumulate, while the position of the population flickers. The  $x$ -axis represents contiguous landscapes which accumulate slight changes from  $\tau$  to  $\tau + 1$ . The arrows mark some of the sudden transitions in the population distribution. (b) Histograms of differences between the state of the population (black) and fitness landscapes (red) in subsequent equilibria  $\tau$  and  $\tau + 1$ . A series of  $2 \cdot 10^6$  consecutive fitness landscapes has been used. While the distribution of changes in the landscape is narrow, that of changes in the population state has a power-law-like tail. Parameters are  $N = 8$ ,  $K = 4$ ,  $f_l = 0.55$ , with an alphabet of two letters and environmental noise  $\epsilon = 0.01$ .

section S3 of the Supplementary Information we calculate how the variations in the fitness and population of a node are distributed in the equilibrium at contiguous environments  $\tau$  and  $\tau + 1$ , following the model presented here. We provide a proof of the qualitative difference between both distributions: the stochastic variable characterising changes of fitness in a single node has a Gaussian, narrow distribution, while the response of the population of that node to changes in fitness can be arbitrarily large. Hence, the distribution corresponding to the latter variable can develop a long, fat tail. The calculations yield:

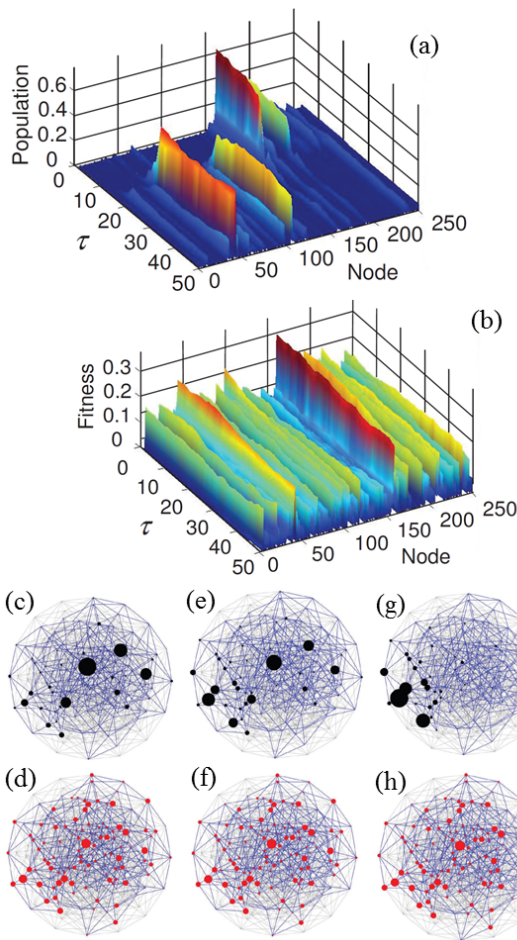
$$\text{Var}(f_i(\tau + 1) - f_i(\tau)) = \frac{\epsilon^2}{3N}, \quad (14)$$

$$\text{Var}((u_1)_i(\tau + 1) - (u_1)_i(\tau)) \propto \frac{\epsilon^2}{N(\lambda_1 - \lambda_2)^2}, \quad (15)$$

that is, the variance of changes in the fitness of each genome  $i$  of length  $N$  is proportional to the squared noise amplitude  $\epsilon^2$  (and thus is small as far as the noise intensity is also small), while that of changes in its population can grow notably because it is inversely proportional to the spectral gap of the transition matrix  $M(\tau)$  (i.e. the difference between the two largest eigenvalues,  $\lambda_1 - \lambda_2$ ) which can be arbitrarily close to zero. This relation with the spectral gap is the key to understand why sudden transitions punctuate amply different stable states of the population while they do not entail a significant change in fitness. The spectral gap will be small in networks that show large heterogeneity, both in the degree distribution or in the weight associated to links and nodes. In the model presented here, this can be achieved only by (i) increasing the value of the lethality coefficient  $f_b$ , which erases nodes and therefore includes heterogeneity in the degree, (ii) increasing the ruggedness of the fitness landscape  $K$ , which introduces heterogeneity in the weight of the nodes, or (iii) increasing the length of the sequences  $N$ , as it is known that large networks show sharper transitions<sup>28,30</sup>. In summary, the mutation-selection equilibrium appears as a complex interplay between the mutation rate, the local fitness of genomes and their correlations, and the global connectivity patterns. Local information is insufficient to predict the collective state of the population, which is summarised in the eigenvector  $\vec{u}_1$ . Also, the growth rate of the population at equilibrium (the effective population fitness given by  $\lambda_1$ ), has a value that does not coincide in general with a specific local or global maximum. That is to say, the sequence of highest fitness is not representative in general of the behaviour of the population.

Finally, we analyse how the genomic population evolves over the space of genomes during a genomic shift. In order to study the phenomenon from this different perspective, Fig. 2 describes in detail a numerical example for a population of sequences evolving in a stochastically varying environment (same parameters as in Fig. 1). The fraction of population of each node -or sequence-  $i$  (given by the  $i$ -element of the eigenvector  $\vec{u}_1$  of the transition matrix  $M(\tau)$ , Fig. 2(a)) and its associated fitness ( $f_i$ , Fig. 2(b)) are plotted for 50 subsequent environments. A sudden shift is observed in  $\tau \simeq 20$ , while no drastic change in fitness values is detected there. In Fig. 2 (c,e,g) and (d,f,h) the population and the fitness of each sequence are plotted over the space of genomes, in three critical situations: just before ( $\tau = 18$ ), during ( $\tau = 20$ ) and after ( $\tau = 22$ ) the tipping point. See the Supplementary Video S1 for an animated plot of this phenomenon.

**Driven fitness landscape.** The second form in which the fitness landscape is varied in this work attends to the fact that changes in environmental conditions are often driven, showing a persistent increase or decrease in value superimposed to stochastic variations<sup>2,8</sup>. In natural systems, gradually changing environments have been shown to induce sudden shifts in population dynamics<sup>31,32</sup>. This situation can be implemented by generating two different fitness



**Figure 2 | Description of a sudden shift in genomic space of a population responding to a stochastically varying environment.** (a) Fraction of population  $(u_i)_i(\tau)$  and (b) fitness of each genome  $f_i(\tau)$  for 50 successive environments. Genomes (or nodes in the space of genomes) are arbitrarily ordered along the  $node$ -axis; successive environments are listed in the  $\tau$ -axis. (c, d) Before the genomic shift occurs ( $\tau = 18$ ), the most populated genome coincides with that of maximal fitness. (e, f) At the transition ( $\tau \approx 20$ ), the population covers a broader region of the space of genomes. After the transition ( $\tau = 22$ ), the population is completely displaced to a new region (g), while changes in the fitness landscape are minor (h). The underlying space of genomes appears in pale grey; dark blue links connect genomes with positive fitness. Size of circles is proportional to sequence population (c, e, g) or to fitness (d, f, h). Parameters as in Fig. 1, and thus the size of the space of genomes is  $A^N = 2^8 = 256$  nodes. See the Supplementary Video S1.

landscapes and interpolating linearly between the corresponding states of fitness for each genome. An additional parameter  $\beta \in [0, 1]$  weights the relative contribution of each landscape.

The methodology is the following. As described in Methods, we generate a starting and an ending fitness landscape,  $\mathcal{L}_{\tau_0}$  and  $\mathcal{L}_{\tau_f}$ , respectively, with the same parameters  $N, K$ , and  $A$ . For better visualisation of the process, and without loss of generality, we relabel the genomes such that those with maximum fitness in either landscape are at the largest Hamming distance. That is to say, genome  $(0, 0, 0, \frac{1}{4})$  is assigned maximum fitness in the starting landscape by fixing  $L_{i_1, j}(\tau_0) = 1 \forall j$ , and sequence  $(1, 1, 1, \frac{1}{4})$  has maximum fitness in the ending landscape by fixing  $L_{i_2, j}(\tau_f) = 1 \forall j$ , where  $i_1$  and  $i_2$  are the rows corresponding to  $K + 1$  zeros and  $K + 1$  ones respectively.

The elements of matrix  $L_\tau$  corresponding to an intermediate landscape  $\mathcal{L}_\tau$  between  $\tau_0$  and  $\tau_f$  become

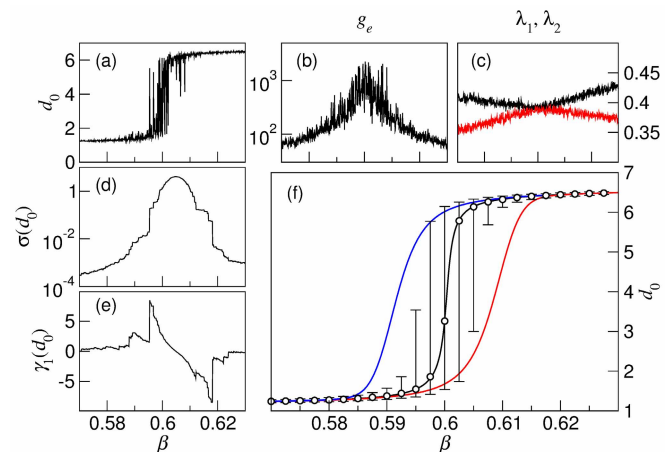
$$(L_\tau)_{ij} = (L_{\tau_0})_{ij} + \beta \left[ (L_{\tau_f})_{ij} - (L_{\tau_0})_{ij} \right] + \epsilon \eta_{ij}, \quad (16)$$

where parameter

$$\beta = \frac{\tau - \tau_0}{\tau_f - \tau_0} \in [0, 1]$$

quantifies the linear transformation of the starting fitness landscape  $\mathcal{L}_{\tau_0}$  (corresponding to  $\beta = 0$ ) into the ending landscape  $\mathcal{L}_{\tau_f}$  ( $\beta = 1$ ),  $\eta_{ij}$  is a random number uniformly distributed in  $[-1, 1]$ , and  $\epsilon$  is the noise intensity. While noise is unnecessary to observe a genomic shift in this situation, it superimposes small fluctuations to the monotonic evolution of the fitness landscape, and allows for the estimation of several important quantities such as early warning signals based on such fluctuations. Once matrix  $L_\tau$  is calculated, the fitness of every genome is obtained following the steps shown in Methods.

Figure 3 shows a numerical genomic shift in a driven and stochastic fitness landscape that evolves as explained above. The abrupt transition occurs at a particular value of  $\beta$  around which stochasticity causes strong flickering, as can be seen in the evolution of the genetic distance  $d_0$  (Eq. (7), Fig. 3(a)). We have chosen the genetic distance to characterise this phenomenon because of its easy experimental measurement through deep-sequencing techniques (see below), but other quantities such as the average fitness of the population  $\bar{F}$  (Eq. (8)) would yield equivalent results. The time to reach mutation-selection equilibrium  $g_e$  grows significantly in the proximity of the transition (Fig. 3(b)). This quantity is larger the smaller the spectral gap  $\lambda_1 - \lambda_2$ , as can be inferred from Eq. (4). Due to stochasticity,  $\lambda_1$  and  $\lambda_2$



**Figure 3 | Early-warning signals and hysteresis in genomic shifts in a driven and stochastic fitness landscape.** (a) Flickering appears around the transition, as measured by the genetic distance  $d_0$  of the population at equilibrium to genome  $\vec{0}$ . (b) Time to equilibrium  $g_e$  dramatically increases close to the transition. (c) Eigenvalues  $\lambda_1$  (black) and  $\lambda_2$  (red) of the transition matrix. The closer they are, the larger  $g_e$ . The spectral gap is typically large in regular networks<sup>38</sup>, while it can be arbitrarily close to zero in sufficiently large, heterogeneous networks. (d) Standard deviation  $\sigma$  and (e) skewness  $\gamma_1$  of  $d_0$  at equilibrium (averaged over  $10^2$  consecutive landscapes). (f) The red and blue curves show the hysteretic behaviour of the genetic distance  $d_0$  of an out-of-equilibrium population subject to regular perturbations each  $G = 1$  generations, for increasing and decreasing values of  $\beta$ . The black curve plots the genetic distances  $d_0$  at equilibrium for fixed  $\beta$  and  $10^3$  realisations of the stochastic perturbations. Open circles represent the median of the distribution of  $d_0$  values; bars indicate the confidence interval 5–95%. Parameter  $\beta$  is sampled at intervals of  $\Delta\beta = 10^{-4}$ ; other parameters are  $N = 8, K = 4, f_i = 0.53, \epsilon = 0.01$  and  $A = 2$ .



fluctuate around average values that become closer as the transition is approached (Fig. 3(c)), but note that the Perron–Frobenius theorem precludes  $\lambda_1 = \lambda_2$  in networks of finite size<sup>30</sup>. For high and low values of  $\beta$ , far from the transition, the population is located in fairly stable and dissimilar states. Close to the transition, however, the stability of the two states is comparable, and they compete to attract the population. This is reflected in the standard deviation of the position of the population calculated as  $\sigma(d_0)$  (Fig. 3(d)), which systematically increases as the tipping point is approached. An additional early-warning signal is skewness  $\gamma_1(d_0)$ , the third standardised moment, a measure of the asymmetry in the fluctuations between the two states (Fig. 3(e)).

The dramatic increase in the time to equilibrium  $g_e$  implies that, close to the transition, populations will not be able to attain high-fitness regions in the presence of too rapid environmental changes. A situation of fast change can be simulated by perturbing the landscape regularly, that is, increasing  $\beta$  every  $G$  generations. The concomitant delay in adaptation will show up when  $G$  is of the order of the time to equilibrium  $g_e$  or smaller. The inability of the population to adapt before the new perturbation arrives (i.e. before the mutation-selection equilibrium is reached) causes hysteresis in its genomic composition, both for increasing and decreasing values of  $\beta$  (red and blue curves respectively in Fig. 3(f), where the genetic distance  $d_0$  of the population was plotted for  $G = 1$ ).

In real cases, data to evaluate standard deviation and skewness of the genetic distance or any other experimental measure over several subsequent mutation-selection equilibria (varying  $\beta$ ), may not be available. Instead, independent populations in similar environments might be at hand. In our model this would correspond to performing different realisations of the noise for each value of parameter  $\beta$ . Open circles in Fig. 3(f) represent the median of the distribution of  $d_0$  values obtained at equilibrium after  $10^3$  realisations of the stochastic signal, and bars indicate the confidence interval, which reveals a strong asymmetry in the distribution. The dispersion of the state of the mutation-selection equilibria so obtained and their asymmetry also offer a clear signal that the transition is approaching.

Finally, note that the NK model yields a rugged landscape, devoid of strictly neutral regions, for any  $K \geq 1$ <sup>16</sup>. In our scenario, it is analogous to an ensemble of quasi-neutral networks where neighbouring nodes are populated thanks to high mutation rates, complex local topology, and fitness correlations. However, analogous transitions are observed in holey-like fitness landscapes<sup>26,27,33</sup>, where large neutral regions are dominant. The early warning signals here used are easily applicable to that situation.

**Robustness of the model.** Most assumptions of our model can be modified to simulate more realistic scenarios without causing qualitative changes in the phenomenology described. A full study of the generalisation of the results here presented is developed in section S2 of the Supplementary Information. We show in Supplementary Figure S1 that increasing the length of the genome, the degree of ruggedness of the landscape (or epistasis), the fraction of lethal mutations and the intensity of environmental changes, as well as decreasing the mutation rate, will increase the frequency and sharpness of transitions. Finally, we have not found major differences when the size of the alphabet  $A$  is varied: Some examples with  $A = 4$  are studied, showing that also for alphabets of 4 letters (corresponding to RNA or DNA) increasing the length of the sequence, the ruggedness of the landscape and the fraction of lethal mutations enhance the intensity of transitions.

## Discussion

Once a transition occurs, it is very difficult for the system to return to its previous state<sup>8</sup>. Therefore, any measure yielding reliable information on the existence of a near tipping point and the concomitant warning signals is extremely valuable. Here we have proved that

abrupt transitions at the genomic level preceded by early-warning signals are a generic phenomenon in constant and slowly driven landscapes affected by even slight stochasticity. We have used a very general model for the evolution of populations on fitness landscapes, and shown that most assumptions can be modified to implement a variety of scenarios without causing qualitative changes in the phenomenology described.

The only essential feature to observe sudden shifts in the genomic composition of a population is the existence of different regions in the landscape separated by fitness barriers, for which a non-negligible fraction of lethal mutations is a sufficient condition. Under environmental changes, the degree of adaptation attained by different genotypic clusters varies and sudden transitions set in. There are at least four features characteristic of some natural systems that actually enhance the abruptness of the transition. First, a higher degree of roughness of the landscape, which accentuates the effect of separating barriers. Second, a lower mutation rate in the population. Finite populations evolving at low mutation rates are less disperse in the space of genomes, even concentrating in a single genome if formed by identical individuals. These populations can explore the vicinity of the fitness landscape only close to the tipping point, and thus are expected to experience stronger hysteresis and more violent genomic shifts. Third, recombination might induce multiple equilibria and thus further separation into distinct genotypic clusters<sup>34</sup>. Fourth, a non-linear relationship between environment and fitness, as indirectly suggested by ecological extinctions caused by smooth environmental changes<sup>35</sup>. In such a case, also the genomic transition would be more severe, and eventually act as the ultimate cause of species extinction.

At larger scales, the controversial possibility of state shifts in the biosphere<sup>8,9</sup> should be preceded by signals unfolding at a hierarchy of levels. At the lowest level, changes in the genomic composition of a population might predate population extinction, which might in turn affect the structure of ecosystems and lead to extinction cascades<sup>36</sup>. These effects could be subsequently transmitted bottom-up. The circle closes when realising that top-down effects might also occur under sudden climatic changes, which might preclude the adaptation of populations, as described. The extent to which the phenomenology presented here has implications at the ecological level and above deserves further exploration.

It is now possible to track the evolution of fast-mutating organisms, such as RNA viruses, through measures as simple as their consensus sequence, which corresponds to the most abundant nucleotide at each position. In controlled environments, rapid fixation of mutations (i.e. changes of nucleotides) in the consensus sequence might indicate that a genomic shift is taking place. Advances in deep sequencing open an interesting avenue to systematically probe the complex molecular structure of populations<sup>37</sup>, since this technique goes beyond consensus sequences to characterize single sequences actually present in the population. Quantitative measures of relative abundances of the different genotypes in a population yield a first approximation to the state eigenvector  $\vec{u}_1$ , and might allow a characterisation of fitness landscapes and genotype network topology at an unprecedented level of detail. Joining measures of the growth rate of populations at equilibrium (corresponding to  $\lambda_1$ ) and estimations of the resilience of populations, now at hand<sup>6</sup> (retrieving  $\lambda_2$ ), to predict the maximum approach of both quantities, should permit a first estimation of the time left before a genomic transition occurs. This could be of the maximum importance, as while early warnings can be used as an indicator of proximity of the tipping point, predicting the actual moment of the transition remains a challenge<sup>2</sup>.

At odds with other systems whose structure can be modified in order to avoid or regulate the appearance of critical transitions, genomic shifts may be unavoidable: the architecture of genotype networks is inherent to the chemical nature of life and its molecular organisation. A much deeper understanding of fitness landscapes



and mutational mechanisms is needed before we can embark in the eventual control of genomic shifts. The artificial enhancement of barriers to adaptation or of the pace of environmental changes, even at the local scale, would be of the highest relevance to achieve that goal.

- Scheffer, M. *et al.* Anticipating critical transitions. *Science* **338**, 344–348 (2012).
- Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
- Pierini, S. Stochastic tipping points in climate dynamics. *Physical Review E* **85**, 027101 (2012).
- Ganopolski, A. & Rahmstorf, S. Abrupt Glacial Climate Changes due to Stochastic Resonance. *Physical Review Letters* **88**, 038501 (2002).
- Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
- Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science* **336**, 1175–1177 (2012).
- Meijer, M. L. *Bio-manipulation in the Netherlands – 15 years of experience.* (Wageningen Univ., Wageningen, The Netherlands, 2000).
- Barnosky, A. D. *et al.* Approaching a state shift in Earth's biosphere. *Nature* **486**, 52–58 (2012).
- Brook, B. W., Curllis, E. C., Perring, M. P., Mackay, A. W. & Blomqvist, L. Does the terrestrial biosphere have planetary tipping points? *Trends in Ecology and Evolution* **28**, 396–401 (2013).
- Eldredge, N. & Gould, S. J. Punctuated equilibria: an alternative to phyletic gradualism. In *Models in Paleobiology*, 82–115 (Freeman, Cooper & Co, USA, 1972).
- Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification - a 10,000 generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences USA* **91**, 6808–6814 (1994).
- Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. Epochal evolution shapes the phylodynamics of inter-pandemic influenza A (H3N2) in humans. *Science* **314**, 1898–1903 (2006).
- Huynen, M. A., Stadler, P. F. & Fontana, W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences USA* **93**, 397–401 (1996).
- van Nimwegen, E. & Crutchfield, J. P. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bulletin of Mathematical Biology* **62**, 799–848 (2000).
- Wilke, C. O. Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution* **55**, 2412–2420 (2001).
- Kauffman, S. A. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* **128**, 11–45 (1987).
- Østman, B. & Adami, C. Predicting evolution and visualizing high-dimensional fitness landscapes. In Engelbrecht, A. & Richter, H. (eds.) *Recent Advances in the Theory and Application of Fitness Landscapes*, Springer Series in Emergence, Complexity, and Computation, to appear (Springer, 2013).
- Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610–618 (2007).
- Lalic, J., Cuevas, J. M. & Elena, S. F. Effect of host species on the distribution of mutational fitness effects for an RNA virus. *PLoS Genetics* **7**, e1002378 (2011).
- Vale, P. F., Choisy, M., Froissart, R. R. S. & Gandon, S. The distribution of mutational fitness effects of phage phiX174 on different hosts. *Evolution* **66**, 3495–3507 (2012).
- Hietpas, R. T., Bank, C., Jensen, J. D. & Bolon, D. N. A. Shifting fitness landscapes in response to altered environments. *Evolution* **67**, 3512–3522 (2013).
- Fisher, K. H. & Hertz, J. A. *Spin glasses* (Cambridge University Press, 1991).
- Kauffman, S. A. & Johnsen, S. Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *Journal of Theoretical Biology* **149**, 467–505 (1991).
- Kauffman, S. A. *The Origins of order: Self-organization and selection in evolution* (Oxford University Press, 1992).
- Newman, M. E. J. & Engelhardt, R. Effects of selective neutrality on the evolution of molecular species. *Proceedings of the Royal Society London B* **265**, 1333–1338 (1998).
- Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology and Evolution* **12**, 307–312 (1997).
- Aguirre, J., Buldú, J. M. & Manrubia, S. C. Evolutionary dynamics on networks of selectively neutral genotypes: Effects of topology and sequence stability. *Physical Review E* **80**, 066112 (2009).
- Aguirre, J., Papo, D. & Buldú, J. M. Successful strategies for competing networks. *Nature Physics* **9**, 230–234 (2013).
- Drake, J. W., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
- Capitán, J. A. & Cuesta, J. A. Catastrophic regime shifts in model ecological communities are true phase transitions. *Journal of Statistical Mechanics* **10**, P10003 (2010).
- van Nes, E. H. & Scheffer, M. Large species shifts triggered by small forces. *The American Naturalist* **164**, 255–266 (2004).
- Harley, C. D. G. & Paine, R. T. Contingencies and compounded rare perturbations dictate sudden distributional shifts during periods of gradual climate change. *Proceedings of the National Academy of Sciences USA* **106**, 11172–11176 (2009).
- Wilke, C. O. Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology* **63**, 715–730 (2001).
- Paixão, T., Bassler, K. & Azevedo, B. R. Emergent speciation by multiple Dobzhansky-Muller incompatibilities. <http://biorxiv.org/content/early/2014/08/21/008268.short> (2014) (Date of access: 21/12/2014).
- Hare, S. R. & Mantua, N. J. Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Prog. Oceanogr.* **47**, 103–145 (2000).
- Estes, J. A. *et al.* Trophic downgrading of planet earth. *Science* **333**, 301–306 (2011).
- Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
- Donetti, L., Neri, F. & Muñoz, M. A. Optimal network topologies: expanders, cages, ramanujan graphs, entangled networks and all that. *Journal of Statistical Mechanics* **2006**, P08007 (2006).

## Acknowledgments

The authors acknowledge financial support from Spanish MICINN (project FIS2011-27569) and from Comunidad de Madrid (MODELICO, S2009/ESP-1691). The authors are indebted to José A. Capitán, José A. Cuesta, Santiago F. Elena, Jaime Iranzo, Carlos Lugo, Ignacio Pascual and Ricard Solé for constructive criticism of the manuscript.

## Author contributions

J.A. and S.M. conceived and designed the project, discussed the results and wrote the manuscript. S.M. coordinated the project. J.A. performed the numerical and analytical work.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Aguirre, J. & Manrubia, S. Tipping points and early warning signals in the genomic composition of populations induced by environmental changes. *Sci. Rep.* **5**, 9664; DOI:10.1038/srep09664 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>