An evolution-based high-fidelity method of epistasis measurement: theory and application to influenza

Gabriele Pedruzzi and Igor M. Rouzine*

Sorbonne Université, Institute de Biologie Paris-Seine

Laboratoire de Biologie Computationelle et Quantitative LCQB

F-75004, Paris, France

Influenza protein sequence data are from https://www.fludb.org.

Short Title: High-fidelity epistasis measurement

Key words: mathematical model, quasi-equilibrium, false-positive, linkage

***Corresponding author:**
Igor Rouzine
igor.rouzine@sorbonne-universite.fr
Evolution and immunology of pathogens
Laboratory of Computational and Quantitative Biology
7238 CNRS-UPMC
Bâtiment C, C311C
Sorbonne Université Campus Pierre et Marie Curie
4 rue Jussieu, Paris 75005
+33 7 69 54 91 28
http://www.ibps.upmc.fr/fr/Recherche/umr-7238/evolution-immunologie-pathogenes

Abstract

Linkage effects in a multi-locus population strongly influence its evolution. The models based on the traveling wave approach enable us to predict the speed of evolution and the statistics of phylogeny. However, predicting the evolution of specific sites and pairs of sites in the multi-locus context remains a mathematical challenge. In particular, the effects of epistasis, the interaction of gene regions contributing to phenotype, is difficult both to predict theoretically and detect experimentally in sequence data. A large number of false interactions arise from stochastic linkage effects and indirect interactions, which mask true interactions. Here we develop a method to filter out false-positive interactions. We start by demonstrating that the averaging of the two-way haplotype frequencies over a multiple independent populations is necessary but not sufficient, because it still leaves high numbers of false interactions. To compensate for this residual stochastic noise, we develop a triple-way haplotype method isolating true interactions. The fidelity of the method is confirmed using simulated genetic sequences evolved with a known epistatic network. The method is then applied to a large database sequences of neurominidase protein of influenza A H1N1 obtained from various geographic locations to infer the epistatic network responsible for the difference between the pre-pandemic virus and the pandemic strain of 2009. These results present a simple and reliable technique to measure site-site interactions from sequence data.

Author's summary

Interaction of genomic sites creating "fitness landscape" is very important for predicting the escape of viruses from drugs and immune response and for passing through fitness valleys. Many efforts have been invested into measuring these interactions from DNA sequence sets. Unfortunately, reproducibility of the results remains low, due partly to a very small fraction of interaction pairs, and

partly to stochastic noise intrinsic for evolution masking true interactions. Here we propose a method based on analysis of genetic sequences at three genomic sites to clean stochastic linkage and apply it to influenza virus sequence data.

Introduction

Almost a century ago, it was realized that the evolution of a population is strongly affected by the fact that the fates of alleles at different loci are linked unless separated by recombination. These *linkage* effects include clonal interference (FISHER 1930; MULLER 1932), background selection, genetic hitchhiking (RICE 2004), enhanced accumulation of deleterious mutations (Muller ratchet) (FELSENSTEIN 1974), and the increase of genetic drift on one locus due to selection at another (HILL AND ROBERTSON 1966). Linkage decreases the speed of adaptation and creates random associations between pairs of mutations occurring on the same branch of the ancestral tree.

These effects have been taken into account in early mathematical models considering two loci (KIMURA 1994) and, more recently, in the traveling wave approach, which describes an arbitrarily large number of connected sites (ROUZINE *et al.* 2003; DESAI AND FISHER 2007; BRUNET *et al.* 2008; ROUZINE *et al.* 2008; HALLATSCHEK 2011; GOOD *et al.* 2012). These models describe the dynamics of fitness classes and include the factors of selection, mutation, random genetic drift and, sometimes, recombination (ROUZINE AND COFFIN 2005; NEHER *et al.* 2010; ROUZINE AND COFFIN 2010). All these models predict a narrow fitness distribution traveling in the fitness space in a direction depending on the initial conditions and parameters. This "traveling wave" consists from the deterministic bulk and the leading stochastic edge, where generation and establishment of rare beneficial mutations limit the adaptation rate. Alternatively, the distribution may move backwards accumulating more and more deleterious alleles (Muller ratchet). These models are able to express, in the general form, important observable quantities in terms of model

parameters, such as the population size, mutation rate, and the distribution of selection coefficients over loci. The observable quantities include the adaptation rate (ROUZINE *et al.* 2003; DESAI AND FISHER 2007; BRUNET *et al.* 2008; ROUZINE *et al.* 2008), Muller ratchet rate (ROUZINE *et al.* 2003; ROUZINE *et al.* 2008), the conditions of full equilibrium (ROUZINE *et al.* 2003; GOYAL *et al.* 2012), fixation probability of an allele, and the most probable selection coefficient (GOOD *et al.* 2012). The same general approach was used to predict the statistical properties of the ancestral tree (BRUNET *et al.* 2007; ROUZINE AND COFFIN 2007; ROUZINE AND COFFIN 2010; WALCZAK *et al.* 2012; NEHER AND HALLATSCHEK 2013).

Despite of all the progress, prediction of the evolution of specific sites in the multi-site context remains an open question. How do allelic frequencies at each site change in time when the system is adapting? Although they are known to follow random trajectories, what can be said about the average allelic frequency of a given site with a given fitness effect of mutation? Also, what can we say about the evolution of site pairs, especially in the presence of epistatic interaction?

Epistasis, defined as co-selection, the interaction of genes and gene regions contributing to phenotype, is an omnipresent phenomenon (CARLBORG *et al.* 2006). Gene interactions are reported to be responsible for a considerable fraction of the organism's genetic inheritance (ZUK *et al.* 2012). They create fitness valleys that the evolutionary path has to cross (WEISSMAN *et al.* 2009). In pathogens, epistasis facilitates the development of drug resistance and immune escape and impedes reversion of resistant mutations (NIJHUIS *et al.* 1999; LEVIN *et al.* 2000; PIANA *et al.* 2002; HANDEL *et al.* 2006; GONZALEZ-ORTEGA *et al.* 2011; WU *et al.* 2018). Most of HIV variation in untreated patients has been argued to arise from mutations gradually compensating the early escape mutations conferring resistance to the immune response (ROUZINE AND COFFIN 1999).

A large number of approaches have been proposed to measure epistasis from genomic data (CORDELL 2009; CHEN *et al.* 2011; UEKI AND CORDELL 2012). The simplest methods are based on pairwise allelic correlations (HILL AND ROBERTSON 1966; BARTON 1995). The problem with all these approaches is that linkage and indirect interactions, create strong inter-site correlations even between non-interacting

locus pairs, and these false-positive pairs are much more numerous than the true epistatic pairs. Stochastic effects are well-recognized as the most serious obstacle to the detection of epistatic effects by any method (WEI *et al.* 2014). In a single asexual population, stochastic linkage overshadows the epistatic footprint, except in a narrow range of times and parameters (PEDRUZZI AND ROUZINE 2019). The same limitation exists for tree-based methods of epistasis measurements (KRYAZHIMSKIY *et al.* 2011; NEVEROV *et al.* 2015).

A method to filter out false links due to indirect interactions has been developed and applied to protein sequence sets obtained from different species (WEIGT *et al.* 2009; COCCO *et al.* 2018). The technique is based on the combination of a covariance method with the estimates of the direct mutual information. The true interactions are inferred from the condition that entropy (defined as the probability of configuration or the degree of disorder) is maximal given the observed frequencies of single sites and site pairs. A similar technique has been used to measure the fitness landscape of Ab binding regions of HIV surface protein gp120 (LOUIE *et al.* 2018). While this technique eliminates indirect interactions in fairly complex network, it is not designed to exclude stochastic linkage effects. The present work is designed to fill this gap.

Recently, we proposed an evolution-based approach to the problem of detection of epistatic effects (PEDRUZZI *et al.* 2018), which we describe here, since it is relevant for our method detailed in *Results*. We simulated the evolution of a pair of loci within a multi-locus genome in an adapting population. We demonstrated the existence of a broad parameter and time range, where each fitness class has sufficient time to assume the most probable state, which changes slowly in time, as the traveling wave shifts in time. This situation will be referred below as "quasi-equilibrium". The time required for the system to arrive there is $t \gg 1/\langle s \rangle$ (PEDRUZZI *et al.* 2018). In quasi-equilibrium, the entropy $S(t)$ (degree of disorder) is at its current maximum and represents a function of fitness $W(t)$, which slowly changes in time during the process of adaptation: $S(t) = F[W(t)]$. The following derivation employs the fact, predicted by all the traveling-wave models cited above, that the distribution of fitness W is narrow, as given by $\mathrm{std}[W] \ll W$. Quasi-equilibrium is

possible due to a very slow motion of the wave, limited by the generation and establishment of new beneficial mutations at the front edge of the wave (ROUZINE *et al.* 2003; DESAI AND FISHER 2007; NEHER *et al.* 2010; HALLATSCHEK 2011). We consider the regime of adaptation, where the population is very far from full equilibrium (steady state), so that we can neglect with deleterious mutation.

Consider a pair of loci (Fig. 1A) and assume that each can have one of two variants, alleles, either beneficial or deleterious allele, denoted 0 or 1, respectively. We denote with s_1, s_2 the respective fitness effects of mutations. There are four possible haplotypes, 00, 01, 10, 11. We assume that the two sites interact epistatically. If both deleterious alleles are present, they partly compensate each other to the degree E, where $0 < E < 1$. Then, the log fitness values of the four haplotypes are given by

$$W_{00} = 0,\ W_{01} = -s_1, W_{10} = -s_2,\ W_{11} = -(s_1 + s_2)(1 - E),$$

respectively. If E is allowed to change below 0 or above 1, we can obtain any type and sign of epistasis: positive, negative, and overcompensation.

The frequency of each haplotype ij in a population, where i and j are 0 or 1, is denoted f_{ij}. As is easy to understand, f_{ij} is proportional to the number of sequence configurations of the rest of the genome, denoted $\exp(S_{rest})$ (grey rectangles in Fig. 1A). Entropy S_{rest} depends on the fitness of the rest of genome equal to $W - W_{ij}$. Hence, the entropy of each haplotype is $S_{rest} = S(W - W_{ij})$. Since the genome is long, W_{ij} is much smaller than W, so we can expand linearly

$$S(W - W_{ij}) \approx S(W) - \beta W_{ij} \tag{1}$$

where $\beta = -dS/dW > 0$. Thus, the frequency of each haplotype f_{ij} is proportional to the corresponding configuration number, $\exp[S(W - W_{ij})]$. Hence, we obtain

$$f_{10} = f_{00} e^{-\beta s_1}$$

$$f_{01} = f_{00}e^{-\beta s_2}$$
$$f_{11} = f_{00}e^{-\beta(s_1+s_2)(1-E)} \quad (2)$$

where β depends on W and, hence, on time, but is the same for all loci (PEDRUZZI *et al.* 2018). This simple derivation neglects interaction of the site pair with the other sites of the genome. This assumption can be lifted and generalized for any interaction network (PEDRUZZI *et al.* 2018). We also do not need to restrict ourselves to the bi-allelic model, but we will keep this assumption, since it is works in the short term. The Boltzman factor in Eq. 2 was derived originally for a small subsystem connecting to a large thermal reservoir. We remind, however, that we are not discussing the state of equilibrium, but adaptation far from equilibrium.

Eq. 2 predicts, in the general form, how the average haplotype frequencies evolve in time. It is tempting to use it for estimating parameters E, s_1, s_1 from the observed haplotype frequencies, and thus find the fitness landscape for the entire genome (at least, if epistasis is pairwise, as we assumed). In reality, these formulae are not readily applicable to real-life sequences. because of the strong random variation of f_{ij} caused by stochastic linkage effects we have discussed. Of course, the problem of strong linkage noise causing poor reproducibility of various methods [see (WEI *et al.* 2014) for review] is not specific for this method of estimating epistasis. Any attempt to measure epistasis in a single population, whether by measuring covariance from measure D', r^2, mutual entropy, UFE (PEDRUZZI *et al.* 2018) or by tree-based methods (NEVEROV *et al.* 2015) faces the same problem.

The epistatic pairs in a single population are not observable in principle, due to the phylogenetic relationship of sequences in population. The problem is that all the sequences in a population closely resemble the common ancestor, which diverges from the origin in a random direction (PEDRUZZI AND ROUZINE 2019). As a result, any measure of co-variance, or even the use of the entire tree, produces only strong noise of random sign. Co-variation due to random linkage completely masks the epistasis signature in a population. The only way to resolve this issue is by

206 averaging over many independent populations with similar parameters under
207 similar conditions.

208 Thus, without sampling multiple populations, it is not possible to infer
209 epistasis in principle, due to the stochastic nature of phylogenetic relation of
210 sequences. This fundamental limitation cannot be resolved by any existing or future
211 method. The contribution of the present work is not to overcome this limitation,
212 which is not possible in principle, but to demonstrate the existence of a large
213 residual stochastic errors left after averaging over a limited ensemble of
214 populations, and to propose a new method to compensate this error. Here we offer
215 a new technique, based on the application of the quasi-equilibrium argument
216 (above) to triple-way haplotypes. We show analytically and by simulation that a
217 triple-way haplotype test represents an effective way to eliminate residual false
218 links, including linkage and indirect interactions. We will demonstrate its high
219 fidelity in a broad parameter range on the mock sequences obtained by computer
220 simulation.

221 After training the method on simulated sequences, we apply this technique to
222 real virus sequences from an adapting population. Influenza virus evolving in a
223 population was demonstrated to obey the traveling wave theory with an effective
224 selection pressure caused by the accumulating memory cells (ROUZINE AND ROZHNOVA
225 2018; YAN *et al.* 2019). Therefore, it is expected to be amenable to our method. Using
226 8000 influenza sequences obtained from various geographic locations sequences,
227 both before and after the pandemic of 2009, we infer the epistatic network in a
228 surface protein, neurominidase, by comparison of the old and new strains. We chose
229 this specific protein, because it affects infectivity and underwent strong changes
230 when it was replaced by a new strain in 2009. It also has been a target for drug
231 therapy.

232 We note that the old strain and the pandemic strain of H1N1 share 80% of
233 homology. This indicates that, at some point in the past, the two strains had a
234 common ancestor, and later they diverged due to accumulation of mutations. Our
235 aim is find out now how the new sequence variant can obtained by changing the old
236 variant by making multiple mutations, including primary and secondary mutations,

without much loss of fitness. Our method does not depend on the actual path of the divergence of the two variants, nor can it infer the order in which these secondary mutations have emerged, but it can infer the resulting network. The same general approach is routinely used for estimating the epistatic networks from the comparison of proteins of strongly divergent species with presumably weak linkage effects (WEIGT *et al.* 2009; COCCO *et al.* 2018). The difference of our approach is that it filters out false-positive links due to linkage, as well as indirect interactions, rather than indirect interactions alone.

RESULTS

Simulation model to generate sequences for the test. We start by simulating the stochastic asexual evolution of a haploid population using a Wright-Fisher process in the presence of the factors of random mutation, random genetic drift, and directed constant selection (Fig. 2b) (*Methods*). We assume two possible alleles at each locus (site), 0 and 1. The binary simplification provides a major reduction in the computational cost and is especially accurate for relatively conserved sequences. We set some pairs of sites to interact epistatically, with a mutual degree of compensation between deleterious alleles chosen to be E = 0.75. It represents an example of strong compensation , 75%. At lesser E, the detection of epistatic links is worse. We consider a simple epistatic network consisting of double arches (Fig. 1B). The choice of the network topology does not affect the detection power of our method. Indeed, as we show below, most false-positive interactions in the short-term evolution are due to stochastic linkage effects rather than indirect epistatic interactions.

First step: averaging over populations. Genome sequences produced by the simulation demonstrate the presence of strong linkage disequlibrium (LD) originating from three sources: direct epistatic interaction, indirect interaction, and stochastic linkage effect. Out first task is to try to detect possible epistatic

interactions from the simulated sequences using pair-wise association analysis (Fig. 2A). As we mentioned in Introduction, their detection is masked by strong stochastic linkage arising from phylogenetic divergence of independent populations (PEDRUZZI AND ROUZINE 2019). To decrease the linkage effect, we calculate pairwise haplotype frequencies, denoted f_{ij} , for each pair of sites and average them over multiple evolutionary-independent populations of the same size. Then, we perform a pairwise association analysis using a correlation measure we termed "Universal Footprint of Epistasis" (UFE), which is a direct estimator of the degree of mutual compensation of two deleterious alleles, E (PEDRUZZI *et al.* 2018) (*Methods*). More traditional LD measures, including D' and Pearson coefficient r^2, have been shown to generate similar stochastic error (PEDRUZZI AND ROUZINE 2019). We select pairs with high correlation, UFE > 0.6. Before we average over populations, for a single population, the raw cluster of inferred pairs is extremely complex and completely hides true epistatic interactions (Fig 2C). A significant reduction of the number of false-positive bonds is obtained by averaging f_{ij} over 200 independent populations (Fig. 2D). However, the fast majority of remaining links are still false-positive, since their number is much higher than the number of actual interactions (Fig. 2B, green arches).

Second step: Triple-site haplotype method. Although much simpler, the resulting network in Fig. 1D is still densely populated by residual linkage links and indirect links, which are as strong as direct interactions. To detect the residual false links, we use a new procedure based on the quasi-equilibrium approximation. We test the validity analytically in *Methods*, as well as below by the direct comparison of the inferred network with the network set in simulation. The procedure is, as follows.

For every potential link between sites i and j detected in Fig. 2D, we recalculate the haplotype frequencies and UFE, by including only the sequences that contain 0 at one of the adjacent connected nodes of the link and denote it UFE_{ij0} (Fig. 2A, bottom). In other words, we calculate the frequencies of three-way haplotypes 110, 100, 010, and 000. Considering different 0-nodes for all connected nodes of the link, we then find the minimum value of UFE_{ij0} over all these nearest-neighbor

299 nodes. The scatter plot in Fig. 2E demonstrates that, for the false pairs, min(UFE_{ij0})
300 is several-fold smaller than UFE_{ij} (red dots in Fig. 2E). For true links (green dots in
301 Fig. 2B,E), the two correlation measures are nearly the same. The reason for this
302 difference is that a 0-node removes a detour path around the link that causes
303 indirect correlations, thus decreasing the false association of the alleles. An analytic
304 demonstration is given in *Methods*. Therefore, we identify and remove false links as
305 those with low min(UFE_{ij0})/UFE. The choice of the threshold is not crucial, as long as
306 we average f_{ij} over, at least, 20 populations for our model parameter choice, so that
307 the two groups, false and true interactions, remain distinct. As a result, for our
308 example, we obtain 100% perfect detection (Fig. 2B). As a bonus, we obtain accurate
309 estimates for the compensation strength, E, because UFE $\approx$ E within 15% accuracy
310 (Fig. 2F, numbers at green links). We can conclude that the triple-haplotype
311 method, for the parameters tested, has a high fidelity with simulated sequences with
312 a known epistatic network.
313
314 **Application to influenza A virus.** After training our method on simulated
315 sequences, we calculate the interaction network for real virus sequences isolated
316 from an evolving population. Our choice is the surface protein of Influenza A virus
317 strain H1N1, Neuraminidase (NA), known to be an important target of drug therapy
318 and immune response. This protein is one of two proteins that control the virus
319 entry into a host cell, the other being Hemaglutinin. Our aim is to identify epistatic
320 network responsible for the difference between the pre-2009 strain and the
321 pandemic strain of 2009. For this aim, we compare the sequences of the first strain
322 to the sequences of the second strain sampled worldwide.
323 We downloaded 8440 sequences of NA from a public database
324 (https://www.fludb.org). They were collected worldwide, from different geographic
325 locations, between years 2005 and 2010, and included both pre-pandemic and post-
326 pandemic strain. All sequences were aligned. Then, we found the consensus
327 (majority) allele for each aminoacid position. To simplify our task, we binarized the
328 sequences by setting each consensus allele to 0 and each non-consensus ("mutant")
329 allele to 1. This simplification is adequate for the aim of detection of interactions,

330 unless several aminoacid variants are present at a site at similar frequencies, which
331 we found to be a rare occurrence.
332 We observed a bimodal distribution of sequences in the mutant allele
333 frequency per genome with two separate maxima of different height. The low-
334 frequency peak was taller. The bimodal distribution reflects the mixture of two
335 strains, the old and the new, with 80% homology. In order to compensate for
336 unequal sampling from the early strain and the late strain, the more abundant
337 sequences with mutation frequency per genome less than a preset value, d_v , were
338 randomly sampled and down-weighted by a coefficient, D_w , ranging from 5% to
339 50%. To obtain the average pairwise haplotype frequencies, f_{ij} , we repeated the
340 resampling multiple times. Then, we followed the procedure described above (Fig.
341 2C-F) to infer the intra-protein network of interactions (Fig. 3).
342
343 **Sensitivity to sampling parameters.** The resulting network is moderately robust
344 to the variation of (d_v, D_w) (Fig. 3A-F). There exists a plateau region of (d_v, D_w) where
345 the total number of links varies weakly (Fig. 3F). We observe that site 248 in NA
346 represents the primary site connected to multiple compensatory sites (Fig. 3D). The
347 dependence of results on (d_v, D_w), which infers between 15 and 22 compensatory
348 sites, probably originates from the unequal presentation of local subpopulations in
349 the database. As a result, stochastic linkage effects average out less than optimally.
350 Below, we choose the network variant shown in Fig 3D as the "golden middle" of the
351 set.
352
353 **Structural interpretation.** The inferred epistatic sites are shown on the three-
354 dimensional protein structure of NA. The active pocket of NA (purple) serves to bind
355 sialic acid on target cell surface (Fig 4A). This result shows that the inferred primary
356 mutation at residue 248 is located near the active pocket. The inferred
357 compensatory mutations (Fig. 3D) helping NA to restore and improve its fitness are
358 located on protein surface inside alpha-helixes, determining the mutual orientation
359 of beta-sheets. Experimentally, mutation 248 is well-known. It was shown to
360 enhance the low-pH stability of NA (TAKAHASHI *et al.* 2013) and is found in all

361 influenza A H1N1 strains after 2009 pandemic, regardless of a geographic location
362 (SEIBERT *et al.* 2012; BYARUGABA *et al.* 2016; OTTE *et al.* 2016).
363
364
365
366 **DISCUSSION**
367
368 In the present work, we propose an efficient evolution-based method to identify the
369 co-variance caused by epistasis from the co-variance caused by stochastic linkage
370 effects and indirect interactions. First, we average the observed haplotype
371 frequencies over independent populations, then we select the links with a high co-
372 variance, and then we apply a tri-way haplotype test for each candidate bond to
373 eliminate the effects of linked sites due to common inheritance.
374 The tri-way haplotype method is justified using a simple analytical model
375 (*Methods*) assuming a quasi-equilibrium state created by a slowly-moving fitness
376 wave (not to be confused with quasi-linkage equilibrium or mutation-selection
377 balance). The existence of quasi-equilibrium has been tested previously by
378 simulation in a broad parameter range (PEDRUZZI *et al.* 2018; BARLUKOVA *et al.* 2020).
379 Intuitively, because population evolves slowly, the distribution of alleles between
380 sites has sufficient time to obtain the most probably, most disordered state given a
381 current fitness.
382 We use a mock sequence set evolved in a Wright-Fisher population with a
383 known epistatic network to demonstrate the high fidelity of the method in a
384 controlled environment. The method eliminates false bonds caused by both indirect
385 interactions and linkage, at least, in the case of a relatively simple epistatic network
386 and moderately diverse populations (< 40% diversity).
387 We applied this technique to identify primary and secondary mutations in
388 influenza A H1N1 neurominidase, responsible for the difference between the pre-
389 2009 and post-2009 variants. We do not address the origins or the history of the
390 strain. Influenza virus has been shown to map to the traveling wave theory (ROUZINE
391 AND ROZHNOVA 2018; YAN *et al.* 2019), which justifies the use of the quasi-equilibrium

392 assumption. Our method infers a single primary site and 15-20 strong multiple
393 compensatory mutations, which number is in the same general range as the number
394 of compensatory mutations observed, for example, in drug-resistant strains of HIV.
395 The inferred primary mutation, 248, has been observed in all influenza A H1N1
396 strains after 2009 pandemic in various geographic locations (SEIBERT *et al.* 2012;
397 BYARUGABA *et al.* 2016; OTTE *et al.* 2016). I was shown to affect virus infectivity
398 (TAKAHASHI *et al.* 2013). Therefore, our method is capable of pinpointing the well-
399 known primary mutation responsible for the emergence of the pandemic strain.

400 As compared to the existing technique of elimination of indirect interactions
401 developed for different animal species, which diverged for millions of years (Weigt
402 et al. 2009; Cocco et al. 2018), our method is designed to be used for recently
403 diverged populations (thousands of generations or less) of the same species with
404 small to moderate diversity. We have tested it for Hamming distance $2f(1-f) <$
405 0.5. Furthermore, our method is capable of eliminating stochastic linkage, less
406 important when comparing different species (Weigt et al. 2009; Cocco et al. 2018).
407 Where both methods can be potentially applied, such as calculating the fitness
408 landscape of HIV Ab-binding regions (LOUIE *et al.* 2018), our method is much faster
409 computationally, because it is local in the genome. Indeed, we can consider one pair
410 of loci at a time and not to worry about simultaneous optimization of $L^2/2$
411 parameters of the full interaction matrix. Also, it helps to avoid the situation when
412 the number of fitting parameter is too large, and the system is over-defined.

413 To illustrate application our method, we averaged haplotype frequencies
414 over influenza H1N1 sequences obtained from a large number of geographic
415 locations. Influenza sequences represent a meta-population with many islands, not a
416 single well-mixed population. The assumption here is that that the averaging over
417 different geographic location allows to obtain, at least, a partial ensemble average.
418 It is unclear to which extent the lack of the full ensemble average affects the
419 estimate, but at least, we are able to compensate the residual linkage errors within
420 that partial ensemble, something that has not been done before.

421 The main limitation of the proposed approach is that it assumes constant
422 directional selection, as opposed to balancing selection or time-dependent selection,

423 such as occurs under changing external conditions or, in the case of pathogens,
424 under the rising immune response. While the specific case of influenza evolving in a
425 population under accumulating immune memory B cells has been shown to map to
426 the case with constant selection (ROUZINE AND ROZHNOVA 2018; YAN *et al.* 2019), the
427 case of virus evolution within an animal under CTL response is less trivial (BATORSKY
428 *et al.* 2014). Both cases require further investigation.

429 To summarize, we proposed a technique to infer the epistatic effect on
430 evolution of locus pairs and tease it out from stochastic linkage effects. We hope that
431 our quasi-equilibrium approach and further development of this technique will
432 prove useful for all researchers interested in finding fitness landscapes of various
433 organisms from their genetic samples. Also, it will be very valuable to compare
434 these data with the result of deep mutational scanning (LEE *et al.* 2018; HOM *et al.*
435 2019), which we hope to do in the future.

436

437

438 **METHODS**

439

440 **Wright-Fisher model for simulation.** We simulate the evolution of a haploid
441 asexual population of N binary sequences. In an individual genome, each locus (site,
442 nucleotide position, aminoacid position) numbered $i = 1, 2, \ldots, L$ is occupied by one
443 of two alleles, either the wild-type allele, denoted $K_i = 0$, or the mutant allele, $K_i =$
444 1. We use a discrete generation scheme in the absence of generation overlap
445 (Wright-Fisher model). The evolutionary factors included in the model are random
446 mutation with rate μL per genome, constant directed selection, and random genetic
447 drift due to random sampling of progeny. Selection includes an epistatic network
448 with a set strength and topology. Recombination is absent. A previous modeling
449 study shows that moderate levels of recombination can enhance epistatic detection
450 (PEDRUZZI AND ROUZINE 2019). Where use the standard model of fitness landscape
451 with pairwise interaction. The logarithm of the average progeny number, W,
452 depends on sequence $[K_i]$ as given by

453

$$W = -\sum_{i=1}^{L} s_i K_i - \sum_{i<j}^{L} s_{ij} K_i K_j \quad (3)$$

$$s_{ij} = E_{ij}(s_i + s_j)T_{ij} \quad (4)$$

Here the selection coefficients s_i and s_j denote the fitness costs of two deleterious mutations that are partially compensated by each other. By definition, E_{ij} is the degree of compensation of deleterious alleles at sites i and j. Values $E = 0$ and 1 represents no epistasis and full compensation, respectively.

In our example in Fig. 2, we consider a haploid population with initial frequency $f_0 = 0.45$, mutation rate $\mu L = 0.07$, fixed selection coefficient $s = 0.1$, $N = 1000$ individuals, and $L = 40$ sites, and fixed epistatic strength $E = 0.75$.

Quasi-equilibrium approximation. In a broad parameter range, an adapting population represents a slowly-moving, narrow peak in fitness space (Rouzine *et al.* 2003; Rouzine and Coffin 2005; Rouzine *et al.* 2008; Rouzine and Coffin 2010; Good *et al.* 2012). Evolution is slow, because the limiting factor is the addition of a rare beneficial mutation established within a highly-fit genetic background (Rouzine *et al.* 2003; Good *et al.* 2012). Because the fitness distribution moves slowly, the entropy (disorder degree) of the mutation distribution over genomes has enough time to reach its current maximum, restricted by the current average fitness of the population. This situation is called "quasi-equilibrium". At each moment, each fitness class has enough time to reach the most probable, most chaotic state given its fitness. Previously, we verified the validity of quasi-equilibrium in a broad range of parameters and initial conditions after time $\sim 1/\langle s \rangle$ (Pedruzzi *et al.* 2018). When the population finally arrives in mutation-selection equilibrium, the presence of deleterious mutation events compensates the natural selection and creates linkage equilibrium. At this point, the method breaks down.

Linkage measure. Eqs. 1 and 2 can be used to estimate epistatic interaction. We will use a binary measure of allelic correlations (Pedruzzi *et al.* 2018)

$$\mathrm{UFE} = 1 - \frac{\log(f_{11}/f_{00})}{\log(f_{01}f_{10}/{f_{00}}^2)} \quad (5)$$

483 where $f_{00}, f_{10}, f_{01} f_{11}$ are the haplotype frequencies averaged over ensemble.

484 UFE performs similarly to more traditionally used measures, such as
485 Lewontin's D' and Pearson correlation coefficient, r^2 (PEDRUZZI AND ROUZINE 2019) or
486 mutual information (WEIGT *et al.* 2009; COCCO *et al.* 2018). As compared to these
487 measures, UFE has the unique advantage of directly measuring the degree of mutual
488 compensation of two alleles E, provided they are epistatically isolated. If the locus
489 pair does not interact with the other sites in the genome, we have UFE $= E$ (see
490 Eqs. 1 and 2). If they are a part of a network, this measure overestimates E. In the
491 main text, we calculate UFE for every pairs of sites (Fig. 2A). We leave only those
492 pairs where UFE exceeds a set threshold of 0.6.

493 **Tri-site test of false bonds.** To test whether a detected correlation for a pair
494 of sites is direct, and not due to linkage or indirect correlation, for each connected
495 pair i, j , we also calculate the three-way measure

496 $$\mathrm{UFE}_{ij0} = 1 - \frac{\log\left(f_{110}/f_{000}\right)}{\log\left(f_{010} f_{100}/{f_{000}}^2\right)} \quad (6)$$

497 where 0 in the third position selects only for the sequences with the consensus
498 allele 0 at a chosen site connected to one site of the tested pair. We consider all
499 possible connected sites as 0-nodes and calculate the minimum value of UFE_{ij0} over
500 all possible 0-nodes. This method cuts the most important indirect correlation by a
501 "detour" for a site pair. It works for both linkage and indirect correlation.

502 **Test of the method using a simple analytic approach.** To illustrate how
503 the above method works on indirect bonds, let us consider a simple analytic model
504 from (PEDRUZZI *et al.* 2018). We assume that some sites carry deleterious alleles with
505 equal selection coefficient $s_i = s_0$ and a fixed epistatic strength $E_{ij} = E$ (Eq. 3) In
506 this simple case, we can fully characterize a genome by the numbers of interacting
507 allelic clusters of different size. Let k_i denote the number of clusters with i nodes
508 and b_i bonds. Then, from Eq. 3, we can express log fitness W as a sum over clusters
509 of different size (Fig. 1B)

510 $$W \equiv -s_0 f_0 L = -s_0 \sum_{i=1}^{i_{max}} k_i (i - 2E b_i) \quad (7)$$

New notation f_0 has the meaning of the effective frequency of uncompensated mutations that would have the same total fitness W. The number of bonds b_i at $i \geq 3$ depends on the topology of epistatic network, while $b_1 = 0, b_2 = 1$ always. In our example with the network of double arches, we have $b_3 = 2$ (Fig. 1B).

As we mentioned above in *Intro* and *Methods*, we assumed the state of quasi-equilibrium (not to be confused with steady state) determined by the current fitness. Therefore, at each moment of time, numbers k_i are determined by the condition that the entropy of the system is maximum given fitness, Eq. 7. Entropy S is defined as the log number of sequence configurations

$$S = \log\left[\prod_{i=1}^{i_{max}} C_{L_i}^{k_i} (n_i)^{k_i}\right] \tag{8}$$

where L_i is the number of all possible locations for a cluster of size i, and n_i is the number of each cluster's configurations (shapes). The values of L_i and n_i depend on the network topology.

Previously, we applied this argument for several topologies to derive the numbers of clusters of different size, as follows (PEDRUZZI *et al.* 2018). We showed, for the one in Fig. 1B, that the numbers of clusters of size $i = 1, 2$ and 3 in, normalized as $f_i = k_i/L$, are related as

$$f_2 = \frac{1}{3} f_1^{2-2E}, \;\; f_3 = \frac{2}{3} f_1^{3-4E} \qquad E < \frac{3}{4} \tag{9}$$

The 1st and 3d site in each triplet in Fig. 1B do not interact directly, but only indirectly through site 2. For these two sites, the haplotype frequencies are

$$f_{11} = 3f_3 + f_1^2, \quad f_{10} = \frac{3}{2} f_2 + f_1 \tag{10}$$

When epistatic interaction is sufficiently strong, as given by the condition $E > 1/2$, triplets dominate over single alleles and their doubles, as given by $f_1 \ll f_2 \ll f_3$ (Eq. 9). In this case, we can approximate haplotype frequencies as

$$f_{11} \approx 2f_1^{3-4E}, \;\; f_{10} \approx \frac{1}{2} f_1^{2-2E}, \qquad \frac{1}{2} < E < \frac{3}{4} \tag{11}$$

536 Using covariance measure $\mathrm{UFE}_{ij} =$ defined in Eq. 3, we obtain $\mathrm{UFE}_{ij} \approx \frac{2E-1}{2-2E}$. We
537 observe that the indirect covariance between sites 1 and 3 is as strong, ~ 1, as for
538 directly interacting sites. For directly interacting sites 1 and 2 (Fig. 1), we
539 previously obtained $f_{11}^{dir} \approx f_1^{3-4E}, f_{10}^{dir} \approx \frac{1}{3} f_1^{2-2E}$, see (PEDRUZZI *et al.* 2018),
540 *Supplement*, Eqs. (3.29) and (3.30).

541 However, if we calculate UFE_{ij0} instead of UFE_{ij} by including only the
542 sequences with majority allele 0 at site 2, we obtain

543 $$f_{101} \approx f_1^2, \;\; f_{100} \approx f_1 \qquad (10)$$

544

545 instead of Eq. 9, and Eq. 6 yields $\mathrm{UFE}_{ij0} = 0$. Thus, the phantom covariance
546 disappears, when we select only sequences with a majority allele inserted between
547 the two sites tested. This result is intuitively clear: by the definition of fitness (Eq.
548 1), only minority alleles interact with each other, while majority alleles form
549 neutral background. The same method turns out to be extremely effective for
550 eliminating false bonds created by linkage (compare Fig. 2B vs 2E).

551

552 **Sequence preparation.** We have applied the three-way test to influenza virus. We
553 performed a multiple progressive alignment for amino acid sequences obtained for
554 a surface protein of Influenza A virus strain H1N1, Neuraminidase (NA), from public
555 database https://www.fludb.org. We focused on NA because of the massive amount
556 of sequence data, because it was a target of drug treatment, and because strong
557 changes in NA were responsible for the higher infectivity of pandemia in 2009.

558 We have used about 8000 sequences found in the cited database for the
559 period 2000-2010 from various geographic locations. Our aim was to understand
560 the differenc between the protein variants before and after pandemic of 2009,
561 which have 80% of mutual homology. We aimed at discovering only the epistatic
562 sub-network related to that difference, and were not interested in any other
563 epistatic interactions in these proteins. For this end, we compared worldwide
564 samples of sequences from the two strains. We randomly sampled similar amounts

565 of sequences from the first and second strains, and re-sampled them several
566 hundred times. We also checked the robustness of the results to exact sampling size
567 (Fig. 2F, *Results*).
568 Pairwise distances between sequences were computed using pairwise
569 alignment with the Gonnet scoring matrix implemented in MATLAB. To calculate the
570 guide tree we used the neighbor-joining method assuming equal variance and
571 independence of evolutionary distance estimates. The obtained consensus served as
572 a universal reference to binarized data sequences. Before applying the detection
573 algorithm, the protein sequences were binarized, by direct comparison of each
574 sequence to the consensus. Each amino-acid residue was set to 0 or 1 for consensus
575 or non–consensus. Although combining all aminoacid variants per site ignores the
576 specific biochemistry of substitutions, this approach greatly reduces the number of
577 haplotype combinations and also increases the sensitivity by effectively increasing
578 the haplotype frequencies.
579 Next, we measured the mutational frequency for each sequence along
580 sequences and for each site across sequences. The subset of low-diversity sequences
581 with allelic frequency below a cut-off d_v was randomly sampled and down-
582 weighted according to a set coefficientm, D_w. Then, we determined the average
583 pairwise and three-way haplotype frequencies for all pairs and triplets of sites, as
584 described above.
585
586 **3D structures**. To characterize the three-dimensional network of epistatic
587 interaction in Fig. 4, we used software package ChimeraX from internet site
588 https://www.rbvi.ucsf.edu/chimerax/.
589
590 **Data availability.** Influenza sequence data are from public database
591 https://www.fludb.org.
592
593 **Acknowledgements.** We thank Martin Weigt and Alessandra Carbone for useful
594 comments. This research has been funded by Agence Nationale de la Recherche
595 grant J16R389 to IMR, http://www.agence-nationale-recherche.fr/.

References

Barlukova, A., G. Pedruzzi and I. M. Rouzine, 2020 The evolutionary origin of the universal distribution of mutation fitness effect https://biorxiv.org/cgi/content/short/867390v1.

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics 140**:** 821-841.

Batorsky, R., R. A. Sergeev and I. M. Rouzine, 2014 The route of HIV escape from immune response targeting multiple sites is determined by the cost-benefit tradeoff of escape mutations. PLoS Comput Biol 10**:** e1003878.

Brunet, E., B. Derrida, A. H. Mueller and S. Munier, 2007 Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization. Physical Review E 76**:** 041104-041101.

Brunet, E., I. M. Rouzine and C. O. Wilke, 2008 The stochastic edge in adaptive evolution. Genetics 179**:** 603-620.

Byarugaba, D. K., B. Erima, M. Millard, H. Kibuuka, L. Lkwago *et al.*, 2016 Whole-genome analysis of influenza A(H1N1)pdm09 viruses isolated in Uganda from 2009 to 2011. Influenza Other Respir Viruses 10**:** 486-492.

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel and L. Andersson, 2006 Epistasis and the release of genetic variation during long-term selection. Nat Genet 38**:** 418-420.

Chen, C. C., H. Schwender, J. Keith, R. Nunkesser, K. Mengersen *et al.*, 2011 Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. IEEE/ACM Trans Comput Biol Bioinform 8**:** 1580-1591.

Cocco, S., C. Feinauer, M. Figliuzzi, R. Monasson and M. Weigt, 2018 Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys 81**:** 032601.

Cordell, H. J., 2009 Detecting gene-gene interactions that underlie human diseases. Nature reviews. Genetics 10**:** 392-404.

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176**:** 1759-1798.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics 78**:** 737-756.

Fisher, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press, 1958, Oxford, United Kingdom.

Gonzalez-Ortega, E., E. Ballana, R. Badia, B. Clotet and J. A. Este, 2011 Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. Antiviral Res 92**:** 479-483.

Good, B. H., I. M. Rouzine, D. J. Balick, O. Hallatschek and M. M. Desai, 2012 Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci U S A 109**:** 4950-4955.

Goyal, S., D. J. Balick, E. R. Jerison, R. A. Neher, B. I. Shraiman *et al.*, 2012 Dynamic mutation-selection balance as an evolutionary attractor. Genetics 191**:** 1309-1319.

641 Hallatschek, O., 2011 The noisy edge of traveling waves. Proc Natl Acad Sci U S A
642 108: 1783-1787.
643 Handel, A., R. R. Regoes and R. Antia, 2006 The role of compensatory mutations in
644 the emergence of drug resistance. PLoS Comput Biol 2: e137.
645 Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial
646 selection. Genet Res 8: 269-294.
647 Hom, N., L. Gentles, J. D. Bloom and K. K. Lee, 2019 Deep Mutational Scan of the
648 Highly Conserved Influenza A Virus M1 Matrix Protein Reveals Substantial
649 Intrinsic Mutational Tolerance. J Virol 93.
650 Kimura, M., 1994 *Population genetics, molecular evolution, and the neutral theory.*
651 *Selected papers.* The University of Chicago Press, Chicago.
652 Kryazhimskiy, S., J. Dushoff, G. A. Bazykin and J. B. Plotkin, 2011 Prevalence of
653 epistasis in the evolution of influenza A surface proteins. PLoS Genet 7:
654 e1001301.
655 Lee, J. M., J. Huddleston, M. B. Doud, K. A. Hooper, N. C. Wu *et al.*, 2018 Deep
656 mutational scanning of hemagglutinin helps predict evolutionary fates of
657 human H3N2 influenza variants. Proc Natl Acad Sci U S A 115: E8276-E8285.
658 Levin, B. R., V. Perrot and N. Walker, 2000 Compensatory mutations, antibiotic
659 resistance and the population genetics of adaptive evolution in bacteria.
660 Genetics 154: 985-997.
661 Louie, R. H. Y., K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty and M. R. McKay,
662 2018 Fitness landscape of the human immunodeficiency virus envelope
663 protein that is targeted by antibodies. Proc Natl Acad Sci U S A 115: E564-
664 E573.
665 Muller, H. J., 1932 Some genetic aspects of sex. Am. Nat. 66: 118-128.
666 Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations.
667 Proc Natl Acad Sci U S A 110: 437-442.
668 Neher, R. A., B. I. Shraiman and D. S. Fisher, 2010 Rate of adaptation in large sexual
669 populations. Genetics 184: 467-481.
670 Neverov, A. D., S. Kryazhimskiy, J. B. Plotkin and G. A. Bazykin, 2015 Coordinated
671 Evolution of Influenza A Surface Proteins. PLoS Genet 11: e1005404.
672 Nijhuis, M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina *et al.*, 1999 Increased
673 fitness of drug resistant HIV-1 protease as a result of acquisition of
674 compensatory mutations during suboptimal therapy. AIDS 13: 2349-2359.
675 Otte, A., A. C. Marriott, C. Dreier, B. Dove, K. Mooren *et al.*, 2016 Evolution of 2009
676 H1N1 influenza viruses during the pandemic correlates with increased viral
677 pathogenicity and transmissibility in the ferret model. Sci Rep 6: 28583.
678 Pedruzzi, G., A. Barlukova and I. M. Rouzine, 2018 Evolutionary footprint of
679 epistasis. PLoS Comput Biol 14: e1006426.
680 Pedruzzi, G., and I. M. Rouzine, 2019 Epistasis detectably alters correlations
681 between genomic sites in a narrow parameter window. PLoS One 14:
682 e0214036.
683 Piana, S., P. Carloni and U. Rothlisberger, 2002 Drug resistance in HIV-1 protease:
684 Flexibility-assisted mechanism of compensatory mutations. Protein Sci 11:
685 2393-2402.

686 Rice, S. H., 2004 *Evolutionary Theory: Mathematical And Conceptual Foundations*.
687 Sinauer Associated.
688 Rouzine, I. M., E. Brunet and C. O. Wilke, 2008 The traveling-wave approach to
689 asexual evolution: Muller's ratchet and speed of adaptation. Theor Popul Biol
690 73**:** 24-46.
691 Rouzine, I. M., and J. M. Coffin, 1999 Search for the mechanism of genetic variation in
692 the pro gene of human immunodeficiency virus. J Virol 73**:** 8167-8178.
693 Rouzine, I. M., and J. M. Coffin, 2005 Evolution of human immunodeficiency virus
694 under selection and weak recombination. Genetics 170**:** 7-18.
695 Rouzine, I. M., and J. M. Coffin, 2007 Highly fit ancestors of a partly sexual haploid
696 population. Theor Popul Biol 71**:** 239-250.
697 Rouzine, I. M., and J. M. Coffin, 2010 Multi-site adaptation in the presence of
698 infrequent recombination. Theor Popul Biol 77**:** 189-204.
699 Rouzine, I. M., and G. Rozhnova, 2018 Antigenic evolution of viruses in host
700 populations. PLoS Pathog 14**:** e1007291.
701 Rouzine, I. M., J. Wakeley and J. M. Coffin, 2003 The solitary wave of asexual
702 evolution. Proc Natl Acad Sci U S A 100**:** 587-592.
703 Seibert, C. W., S. Rahmat, F. Krammer, P. Palese and N. M. Bouvier, 2012 Efficient
704 transmission of pandemic H1N1 influenza viruses with high-level oseltamivir
705 resistance. J Virol 86**:** 5386-5389.
706 Takahashi, T., J. Song, T. Suzuki and Y. Kawaoka, 2013 Mutations in NA that induced
707 low pH-stability and enhanced the replication of pandemic (H1N1) 2009
708 influenza A virus at an early stage of the pandemic. PLoS One 8**:** e64439.
709 Ueki, M., and H. J. Cordell, 2012 Improved statistics for genome-wide interaction
710 analysis. PLoS Genet 8**:** e1002625.
711 Walczak, A. M., L. E. Nicolaisen, J. B. Plotkin and M. M. Desai, 2012 The structure of
712 genealogies in the presence of purifying selection: a fitness-class coalescent.
713 Genetics 190**:** 753-779.
714 Wei, W.-H., G. Hemani and C. S. Haley, 2014 Detecting epistasis in human complex
715 traits. Nature Reviews Genetics 15**:** 722.
716 Weigt, M., R. A. White, H. Szurmant, J. A. Hoch and T. Hwa, 2009 Identification of
717 direct residue contacts in protein-protein interaction by message passing.
718 Proc Natl Acad Sci U S A 106**:** 67-72.
719 Weissman, D. B., M. M. Desai, D. S. Fisher and M. W. Feldman, 2009 The rate at which
720 asexual populations cross fitness valleys. Theor Popul Biol 75**:** 286-300.
721 Wu, N. C., A. J. Thompson, J. Xie, C. W. Lin, C. M. Nycholat *et al.*, 2018 A complex
722 epistatic network limits the mutational reversibility in the influenza
723 hemagglutinin receptor-binding site. Nat Commun 9**:** 1264.
724 Yan, L., R. A. Neher and B. I. Shraiman, 2019 Phylodynamic theory of persistence,
725 extinction and speciation of rapidly adapting pathogens. Elife 8.
726 Zuk, O., E. Hechter, S. R. Sunyaev and E. S. Lander, 2012 The mystery of missing
727 heritability: Genetic interactions create phantom heritability. Proc Natl Acad
728 Sci U S A 109**:** 1193-1198.
729

730

731

732

733 **FIGURES AND CAPTIONS**

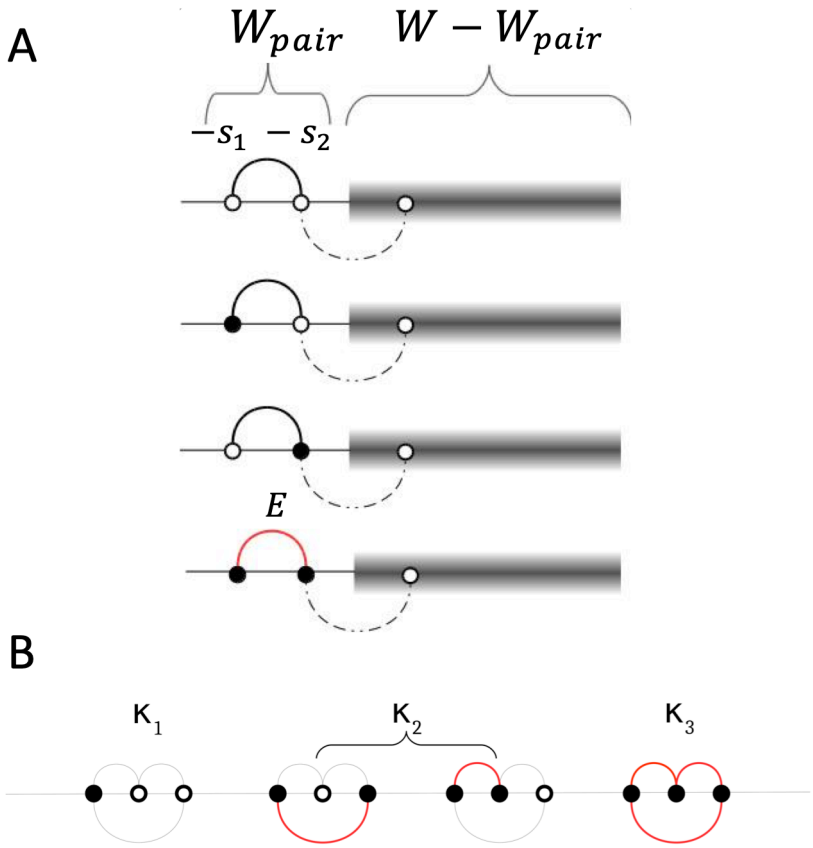

Fig. 1. (A) A pair of interacting sites in a long genome. Black and white circles are deleterious and favorable alleles, respectively, arches show direct (potential) epistatic interactions, red arches are actual interactions between deleterious alleles. Dashed line: neglected interaction. Grey box: the rest of the genome. (B) A simplest epistatic network illustrating the effect of indirect interactions.. Numbers k_i denote the numbers of clusters with i connected sites in a genome.

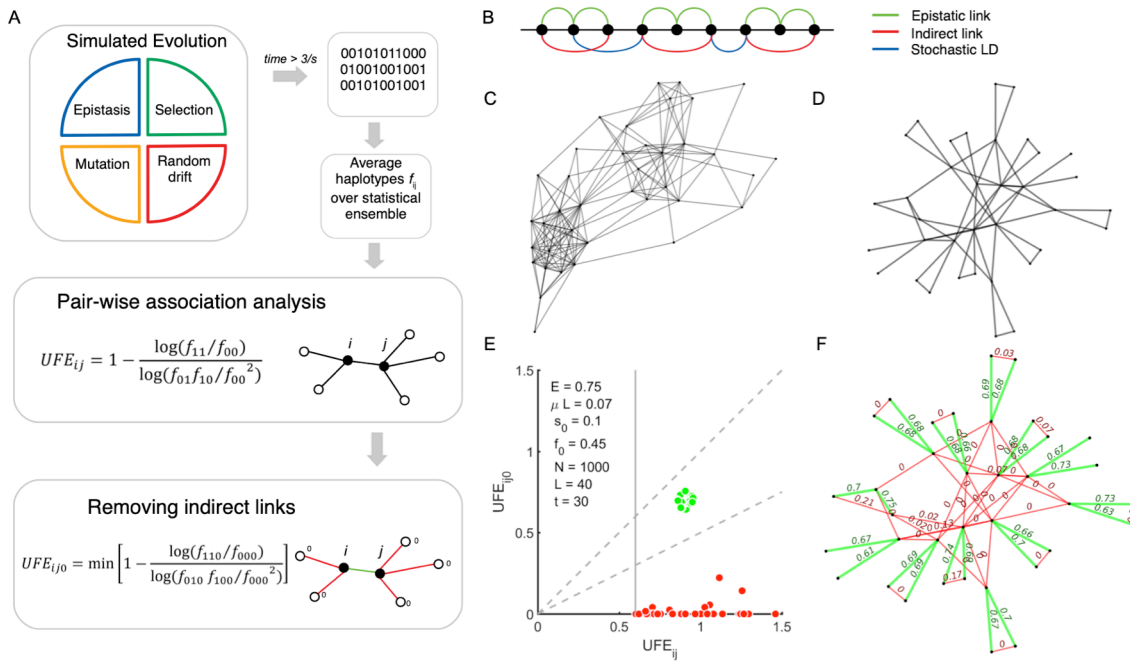

Fig. 2 Schematic diagram of the method and its training on simulated sequence data. A. The computer model of asexual evolution includes the factors of random mutation, selection, epistasis, and random genetic drift. Pairwise haplotype frequencies f_{ij} are averaged over an ensemble of independent simulation runs (populations). The strength of interactions UFE_{ij} is calculated. The indirect links and the residual stochastic LD are excluded by using triple-site haplotype frequencies, UFE_{ij0}. **B.** Pre-set epistatic network for 50 sites. Real epistatic links are shown by green lines. The resulting indirect links are red lines. Some examples of stochastic linkage bonds are shown by blue lines. **C-D.** The network of strong (UFE > 0.5) candidate epistatic interactions predicted (**C**) from a single population and (**D**) after averaging over 200 populations. **E.** Scatter plot of 3-locus haplotype min(UFE_{ij0}) shown against UFE_{ij} for the pairs identified in (**D**). Dashed sector (green): Predicted direct interactions. **F.** Predicted network accurately recapitulates the pre-set epistatic network. Parameters: initial allele frequency $f_0 = 0.45$, mutation rate per genome $\mu L = 0.07$, fixed selection coefficient $s = 0.1$, $N = 1000$, $L = 40$, epistatic strength $E = 0.75$.

734

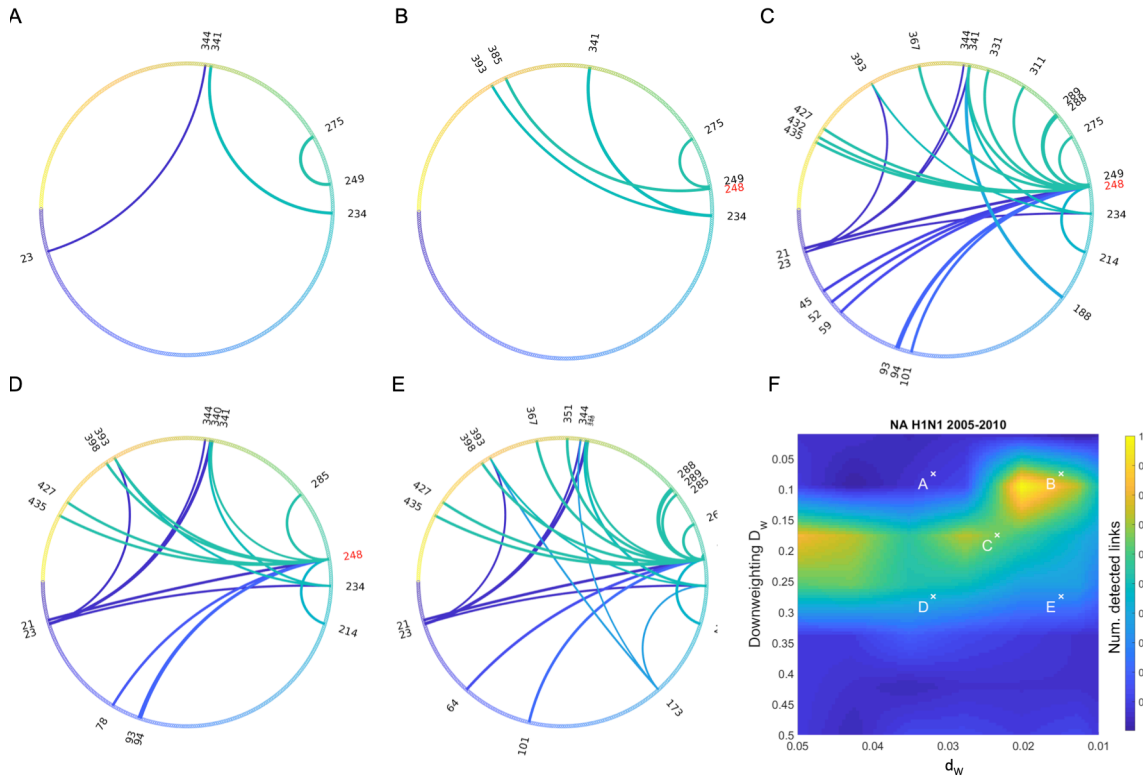

Fig. 3 Epistatic network predicted from sequence data on surface protein sequences of Influenza A H1N1 obtained in 2005 and 2010.

The circular diagrams show the network of interaction for neuraminidase. Sequences with homology to the consensus less than d_v are randomly re-sampled many times, with their number downweighted by coefficient D_w. (**F**) 2D heatmap showing the total number of links as a function of d_v (X-axis) and D_w (Y-axis). Different versions of wheels in **A-E** correspond to different choices of D_w and d_v shown by crosses in **F.**

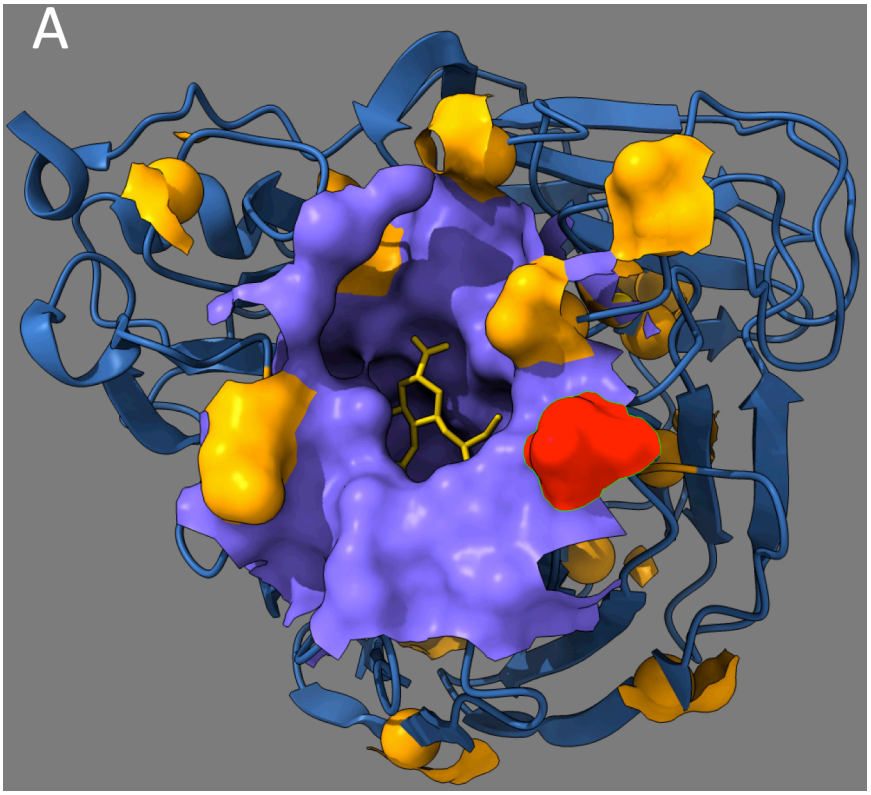

Fig. 4. Structural location of the predicted epistasis network for the neuraminidase of influenza virus. The figure shows the three-dimensional structure of Influenza A H1N1 neuraminidase (PDB ID code 4QVZ). Colored spheres represent predicted epistatic residues from Fig. 3D. Red sphere: Predicted primary mutation (residue 248 in Fig 3D). Orange spheres: Compensatory residues from Fig. 3D.