Physical Biology

OFFICIAL JOURNAL OF
Sociedad de Biofísicos Latino Americanos

PAPER • OPEN ACCESS

DCA for genome-wide epistasis analysis: the statistical genetics perspective

To cite this article: Chen-Yi Gao *et al* 2019 *Phys. Biol.* **16** 026002

Physical Biology

OPEN ACCESS

RECEIVED
11 September 2018

REVISED
4 December 2018

ACCEPTED FOR PUBLICATION
3 January 2019

PUBLISHED
29 January 2019

PAPER

DCA for genome-wide epistasis analysis: the statistical genetics perspective

Chen-Yi Gao[1,2], Fabio Cecconi[3], Angelo Vulpiani[3,4], Hai-Jun Zhou[1,2] and Erik Aurell[5,6,7]

[1] Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[2] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[3] CNR-ISC and Dipartimento di Fisica, Sapienza Universitá di Roma, p.le A. Moro 2, 00185 Roma, Italy
[4] Centro Interdisciplinare B. Segre, Accademia dei Lincei, Roma, Italy
[5] Department of Computational Biology, KTH-Royal Institute of Technology, SE-10044 Stockholm, Sweden
[6] Departments of Applied Physics and Computer Science, Aalto University, FIN-00076 Aalto, Finland
[7] Laboratoire de Physico-Chimie Théorique—UMR CNRS Gulliver 7083, PSL Research University, ESPCI, 10 rue Vauquelin, F-75231 Paris, France

E-mail: gaocy@itp.ac.cn, fabio.cecconi@cnr.it, angelo.vulpiani@roma1.infn.it, zhouhj@itp.ac.cn and eaurell@kth.se

Keywords: direct coupling analysis, genome-scale, quasi-linkage equilibrium

Abstract

Direct coupling analysis (DCA) is a now widely used method to leverage statistical information from many similar biological systems to draw meaningful conclusions on each system separately. DCA has been applied with great success to sequences of homologous proteins, and also more recently to whole-genome population-wide sequencing data. We here argue that the use of DCA on the genome scale is contingent on fundamental issues of population genetics. DCA can be expected to yield meaningful results when a population is in the quasi-linkage equilibrium (QLE) phase studied by Kimura and others, but not, for instance, in a phase of clonal competition. We discuss how the exponential (Potts model) distributions emerge in QLE, and compare couplings to correlations obtained in a study of about 3000 genomes of the human pathogen *Streptococcus pneumoniae*.

1. Introduction

Direct coupling analysis (DCA) is a collective term for a number of related techniques to learn the parameters in Ising/Potts models from data and to use these inferred parameters in biological data analysis [1]. DCA has led to a breakthrough in identifying epistatically linked sites in proteins from protein sequence data [2–6], which in turn has been used to predict spatial contacts from the sequence data [7–9]. DCA has also been used to identify nucleotide–nucleotide contacts of RNAs [10], multiple-scale protein–protein interactions [11, 12], amino acid–nucleotide interaction in RNA-protein complexes [13] and synergistic effects not necessarily related to spatial contacts [14–17], particularly in describing HIV and its interaction with the host immune system [18–20].

Skwark *et al* applied a version of DCA to whole-genome sequencing data of a population of *Streptoccoccus pneumoniae* [21], and were able to retrieve interactions between members of the penicillin-binding protein (PBP) family of proteins, as well as other predictions. *S. pneumoniae* (pneumococcus) is an important human pathogen where resistance to antibiotics in the β-lactam family of compounds are associated to alterations in their target enzymes, which are the PBPs [22]. Further results were recently given in [23] showing robustness by using sequencing data from a pneumococcal population from another continent, and identifying a novel seasonal phenotype signal. Three of the authors of the current article additionally recently showed that DCA analysis on the bacterial genome scale does not need supercomputing resources, but can be carried out in a reasonable time (hours) on a standard desktop computer [24].

These advances raise the question why DCA works at all, and if one can identify from the outset when that is the case. One of us has argued that 'max-entropy' reasoning often evoked in the DCA literature is not pertinent to this issue [25], for an opposite view see [26]. Regardless of one's view on DCA in general, we will here argue that at least for genome-scale data the answer to why it works lies in a very different direction. We will show that the *quasi-linkage equilibrium* (QLE) of Kimura [27–29], as extended by Neher and Shraiman to statistical genetics on the genome scale

[30, 31], provides a natural and rational basis for DCA. According to this theory a population evolving with sufficient amount of exchange of genetic material (recombination, or any form of sex) will settle down to a dynamic equilibrium where the distribution of genotypes is of the form assumed by DCA. In the opposite case of little exchange of genetic material (little sex) the distribution over genotypes is different and dominated by clones, identical or very similar individuals descended from common ancestors. In such a setting straight-forward application of DCA is not an appropriate approach, and is likely to yield nonsense results.

We will also discuss the inference task of DCA in the context of QLE as realistically applied to biological data. We will first show that DCA can give a much sparser representation of the data than correlations (covariances). This is in line with the intended meaning of the acronym DCA: the parameters in a Potts or an Ising model can be considered 'direct couplings', and while these are typically reflected in correlations (covariances), the latter also includes combined effects, or 'indirect couplings'. Second, the authors of [30, 31] assumed that a genotype can be described by a Boolean vector i.e. a string of 0s and 1s. This is almost never the case for population-wide whole-genome sequencing data due to varying gene content, which have to be represented as gaps. We have therefore generalized the theory to categorical data and to a model of bacterial recombination. Third, as surveyed in [1], DCA as a methodology has matured considerably over the last decade. For the mathematical task of inferring parameters in a Potts or Ising model from data which was generated from such a model, the small-interaction expansion (SIE) used in [30, 31] is inferior to many other inference methods. We will show that it is also inferior when applied to real data in the sense of yielding much less sparse results, and would also have specific problems when applied to simulation data. For convenience we include a derivation of SIE as applied to categorical data in appendix A. A conclusion of this work is hence that when QLE is combined with DCA on the genome scale, it should be combined with one of the modern and more powerful versions of DCA.

The paper is organized as follows. Sections 2–5 reformulate the theory of [31] in a way suitable to our presentation and for categorical data. Section 2 hence contains a non-mathematical overview, while sections 3 and 4 contain the specific changes needed for categorical data and our model of bacterial recombination. In section 2 we also define what we mean by 'statistical genetics', and give further background references. Section 5 formulates the dynamics of Potts model parameters in QLE phase, which is a central result of the theory. Section 6 presents results for real sequence data and section 7 for simulation data. Section 8 contains discussion and outlook for future work. Technical details such as a derivation of SIE for categorical data (referred to above), a detailed derivation of the central result in (25), and sequence and code availability are given in appendices.

2. Statistical genetics and quasi-linkage equilibrium

Statistical genetics is the term advanced in [31] to emphasize the formal similarities between the dynamics of genomes in a population and entities (spins) in statistical physics. Such analogies have in fact a long history: Fokker–Planck equations were introduced to describe the change of probability distributions over genotypes by Fisher [32, 33], and Kolmogorov [34]; for more recent physical perspectives see [35] and [36], and references therein. Modern notions of non-equilibrium statistical physics such as entropy production and fluctuation relations have also been shown to have analogies in the theory of biological fitness [37]. Central to the discussion in [30] and [31] is the inclusion of recombination, formally similar to a collision operator, which therefore leads beyond the level of linear models (Fokker–Planck equations). We will in the following use the term statistical genetics in this more restricted sense to emphasize effects of recombination (analogous to non-zero collision rates and the dynamics of non-ideal gases) and present key concepts and results in a mostly non-mathematical manner.

The driving forces of evolution are assumed to be genetic drift, mutations, recombination, and fitness variations. The first refers to the element of chance; in a finite population it is not certain which genotypes will reproduce and leave descendants in later generations. In an infinite population, the latter three are deterministic, describing the expected success or failure of different genotypes. Mutations are hence random genome changes described by mean rates.

Recombination (or sex) is the mixing of genetic material between different individuals. In diploid organisms every individual inherits half of its genetic material from the mother, and half from the father. This material is also before that mixed up in the process called cross-over so that a chromosome of the child inherited from one parent typically consists of segments alternately taken from the two chromosomes of that parent. By sequencing the parents and children in a single family the per generation mutation rate and number of cross-over segments in human has been measured to be about 30 and 100 [38], numbers that are in line with previous estimates. By this measure recombination is hence in human about three times faster than mutations. In bacteria recombination happens by transformation (ability to take up DNA from the surroundings), transduction (transfer of genetic material by the intermediary of viruses), and conjugation (direct transfer of DNA from a donor to a recipient). The ratio of recombination to mutations differ greatly between different bacterial species and can also differ between different strains and differ-

ent environments of the same species. In this work we use data from *S. pneumoniae* where this ratio has been estimated from less than one to over forty, but with an average close to nine [39]. Similarly to the analysis in [31] we will for the most part here assume that recombination is a faster and stronger effect than mutations.

Fitness means in statistical genetics a propensity for a given genotype to propagate its genomic material to the next generation. Like mutation and recombination fitness is hence here a rate, measured in units $(\text{time})^{-1}$. Fitness variations refer to the variations of these rates. Consider then the effects of recombination and fitness on correlated variations in a population, ignoring mutations and genetic drift. The correlation between alleles α and β at loci i and j is $M_{ij}(\alpha,\beta) = f_{ij}(\alpha,\beta) - f_i(\alpha) f_j(\beta)$ where $f_i(\alpha)$ is the frequency of allele α at locus i and similarly $f_j(\beta)$, and where $f_{ij}(\alpha,\beta)$ is the frequency of simultaneously finding alleles α and β at loci i and j. If there is recombination between i and j but are no fitness variations at all, then it is trivial to see that $M_{ij}(\alpha,\beta)$ must decay to zero. This state is called *linkage equilibrium* (LE).

If now instead fitness variations are small but non-zero, then non-zero correlations may persist. We will assume that the fitness of genotype $\mathbf{g}$ which carries allele g_i on locus i depends on single-locus variations and pair-wise co-variations, that is

$$F(\mathbf{g}) = F_0 + \sum_i F_i(g_i) + \sum_{ij} F_{ij}(g_i, g_j). \qquad (1)$$

If so, the first central result of statistical genetics is that when recombination is sufficiently strong, the distribution of genotypes will have the form

$$P(\mathbf{g}) = \frac{1}{Z} \exp\left(\sum_i h_i(g_i) + \sum_{ij} J_{ij}(g_i, g_j) \right). \qquad (2)$$

The above distribution is also the Gibbs–Boltzmann distribution over variables $\mathbf{g}$ with energy terms h_i and J_{ij}, and where Z (the partition function) is the normalization. The second central result is that

$$J_{ij}(\alpha,\beta) = \frac{F_{ij}(\alpha,\beta)}{r c_{ij}} \qquad (3)$$

where r is an overall recombination rate and c_{ij} is the probability that alleles at loci i and j are inherited from different parents. Recombination in real organisms will typically have structure and around 'recombination hotspots' (positions where recombination is more likely), c_{ij} will increase to its asymptotic value $\frac{1}{2}$ at a faster rate. A method to infer recombination hotspots in bacterial genomes was discussed in [40], and the issue was also discussed in the first publication on the 'Maela' data set [41]. For the most part we will in the following assume that c_{ij} equals $\frac{1}{2}$, appropriate if recombination is sufficiently strong and loci i and j are sufficiently far apart on the genome. Note that the right-hand side of (3) is the ratio of two rates, and therefore dimension-less. For the distribution (2) the parameters $J_{ij}(\alpha,\beta)$ carry the same information as the correlations $M_{ij}(\alpha,\beta)$ but in a de-convoluted or 'direct' manner.

Inferring $J_{ij}(\alpha,\beta)$ from data is what the methods known as DCA achieve [1]. From (3) this gives fitness parameters $F_{ij}(\alpha,\beta)$ up to a proportionality (the overall rate r), and for pairs of loci sufficiently far apart on the genome so that c_{ij} is approximately constant. Recombination does not change single-locus frequencies; in a stationary state parameters $h_i(\alpha)$ instead result from a dynamic equilibrium between fitness and mutations. In the absence of mutations QLE in an infinite population is in fact only a long-lived transient while the $h_i(\alpha)$ change slowly in time as the population drifts towards fixation. Non-zero mutations are therefore necessary to maintain the quasi-linkage equilibrium state itself, but are not necessary to describe the properties of that state that we are mainly interested in (the J_{ij}s). In a finite population both $h_i(\alpha)$ and $J_{ij}(\alpha,\beta)$ also fluctuate in time, and the prediction (3) does not apply directly. All these aspects have to be taken into account when applying DCA techniques to analyze a QLE phase.

The third central prediction of statistical genetics is that when fitness variations still have the form (1) but are not small compared to recombination, then the distributions will not be of the form (2). In that phase, in [30, 31] called *clonal competition* (CC), the distribution is instead better described as

$$P(\mathbf{g}) = \sum_c \mu_c P_c(\mathbf{g}) \qquad (4)$$

where the sum goes over clones, μ_c is the weight of clone c, and $P_c(\mathbf{g})$ is some distribution peaked around clone center $\mathbf{g}_c$. Statistical genetics hence predicts a parameter-dependent transition between the two canonical distribution families in high-dimensional statistics, namely the exponential model (2) and the mixture model (4). A further difference between QLE and CC is that in QLE the joint distribution over more than one genotype approximately factorizes, $P(\mathbf{g}_1, \ldots, \mathbf{g}_N) \approx P(\mathbf{g}_1) \cdots P(\mathbf{g}_N)$. In CC phase this is not so; genomes related by descent do not vary independently. The structure of the CC phase was studied in [42], but the problem of inferring fitness parameters in that phase has to our knowledge not been addressed. Although interesting we will have to leave that to future work. We note that even if each of clone distributions $P_c(\mathbf{g})$ is a Potts model distribution, the total distribution $P(\mathbf{g})$ is then a mixture of Potts distribution for which the inference task is computationally considerably more difficult. The only approach we are aware of [43] would likely be difficult to apply on the genome scale.

3. Statistical genetics for categorical data

In this section we summarize statistical genetics as formulated in [30, 31] in a more technical manner, and generalize the theory to categorical data, i.e. to when

there can be more than two alleles per locus. Let there be A_i alleles at locus i and let the allele be indicated by a variable l_i that takes values $1, 2, \ldots, A_i$. The frequency of allele α at locus i is

$$f_i(\alpha) = \langle \mathbf{1}_{l_i,\alpha} \rangle, \tag{5}$$

where $\mathbf{1}_{i,j}$ is the Kronecker delta. These quantities satisfy $\sum_{\alpha=1}^{A_i} f_i(\alpha) = 1$. The covariance matrix between loci i and j is

$$M_{ij}(\alpha,\beta) = \langle \mathbf{1}_{l_i,\alpha} \mathbf{1}_{l_j,\beta} \rangle - f_i(\alpha) f_j(\beta). \tag{6}$$

These quantities satisfy $\sum_{\alpha=1}^{A_i} M_{ij}(\alpha,\beta) = \sum_{\beta=1}^{A_j} M_{ij}(\alpha,\beta) = 0$. A non-zero norm of a covariance matrix (scalar value) is termed linkage disequilibrium (LD) in population genetics [46]; here we deviate from this conventional notation as we are also interested in the individual matrix elements, as well as the formally similar variance matrix at one locus. This variance matrix is

$$M_{ii}(\alpha,\beta) = \mathbf{1}_{\alpha,\beta} f_i(\alpha) - f_i(\alpha) f_j(\beta) \tag{7}$$

and satisfies $\sum_{\alpha=1}^{A_i} M_{ii}(\alpha,\beta) = \sum_{\beta=1}^{A_j} M_{ii}(\alpha,\beta) = 0$.

Statistical genetics are evolution equations for the distributions over genotypes

$$\frac{d}{dt} P(\mathbf{g}) = \frac{d}{dt}|_{\text{mut}} P(\mathbf{g}) + \frac{d}{dt}|_{\text{fitness}} P(\mathbf{g}) + \frac{d}{dt}|_{\text{recomb}} P(\mathbf{g}) \tag{8}$$

where the three terms on the right-hand side represent the changes due to mutations, fitness variations and recombination. The mechanisms of mutations and fitness are classical in population genetics, and referred to as Wright–Fisher or Wright–Fisher-like models [36].

Single-locus mutations are hence modelled by matrices $\mu^{(i)}_{\alpha,\beta}$ which give the rate at which allele α on locus i changes to allele β. Let $F^{(i)}_{\alpha,\beta}$ be the operator which if the allele at locus i is α changes it to β, and otherwise does nothing. In the dynamic equation for probability mutations hence enter as

$$\frac{d}{dt}|_{\text{mut}} P(\mathbf{g}) = \sum_i \sum_{\alpha\beta} \mathbf{1}_{g_i,\alpha} \left(\mu^{(i)}_{\beta,\alpha} P\left(F^{(i)}_{\alpha,\beta}\mathbf{g}\right) - \mu^{(i)}_{\alpha,\beta} P(\mathbf{g}) \right). \tag{9}$$

This gives contributions to the dynamic equations for the frequencies and correlations as

$$\frac{d}{dt}|_{\text{mut}} f_i(\alpha) = \sum_\gamma \mu^{(i)}_{\gamma,\alpha} f_i(\gamma) - \sum_\gamma \mu^{(i)}_{\alpha,\gamma} f_i(\alpha) \tag{10}$$

$$\begin{aligned}\frac{d}{dt}|_{\text{mut}} M_{ij}(\alpha,\beta) = &\sum_\gamma \mu^{(i)}_{\gamma,\alpha} M_{ij}(\gamma,\beta) - \sum_\gamma \mu^{(i)}_{\alpha,\gamma} M_{ij}(\alpha,\beta) \\ &+ \sum_\delta \mu^{(j)}_{\delta,\beta} M_{ij}(\alpha,\delta) - \sum_\delta \mu^{(j)}_{\beta,\delta} M_{ij}(\alpha,\beta).\end{aligned} \tag{11}$$

In the simulations reported below all transition rates $\mu^{(i)}_{\alpha,\beta}$ are the same. As discussed previously it is often a reasonable assumption to take mutations a weaker effect than fitness variations and recombination.

Fitness variations, such as (1), act on the distributions over genotypes as

$$\frac{d}{dt}|_{\text{fitness}} P(\mathbf{g}) = (F(\mathbf{g}) - \langle F \rangle) P(\mathbf{g}) \tag{12}$$

where $\langle F \rangle = \sum_{\mathbf{g}} F(\mathbf{g}) P(\mathbf{g})$ is the instantaneous average of fitness over the population.

Potts models were introduced above in (2). As written that model over-parametrized since the same distribution is found by shifting all $h_i(\alpha)$ by a constant c_i or all $J_{i,j}(\alpha,\beta)$ by a vector $c_{ij}(\beta)$, or a vector $c_{ij}(\alpha)$. In the DCA literature it is customary to go to the Ising gauge [2, 47] given by

$$\sum_\alpha h_i(\alpha) = \sum_\alpha J_{i,j}(\alpha,\beta) = \sum_\beta J_{i,j}(\alpha,\beta) = 0. \tag{13}$$

The fitness function (1) also has this kind of invariance, and we will also use the Ising gauge for this quantity.

4. Bacterial recombination in statistical genetics

Recombination (or sex) takes many different forms depending on if the organism is haploid or diploid and the type of recombination. The mechanism formulated in [30, 31] is specifically for sexual reproduction in haploid yeast, where two parents each produce a mating body (copy of parent genome), and these two mating bodies merge and produce one new genome while the other half of the genetic material of the two mating bodies is discarded. As closer to our data we consider instead a form of bacterial recombination, for which however the evolution essentially turns out to be the same, modulo a *Stosszahlansatz*.

Recombination is thus (we assume) distinguished by two genomes merging and forming two new genomes. This does not directly model conjugation where one bacterium gives genetic material to the other, but can model transformation and transduction over time where material can go both ways. In *S. pneumoniae* recombination happens by transformation and homologous recombination. In an elementary step two genotypes are hence lost (the parents) and two genotypes are gained (the offspring). Let $E_{\mathbf{g}_1,\mathbf{g}_2\to\mathbf{g}'_1,\mathbf{g}'_2}$ be the event that two individuals with genotypes $\mathbf{g}_1$ and $\mathbf{g}_2$ recombine and give two individuals $\mathbf{g}'_1$ and $\mathbf{g}'_2$. To describe the kinetics of the individual process we assume that recombination between the two parents happen with rate $rQ(\mathbf{g}_1,\mathbf{g}_2)$ where r an overall rate of recombination and $Q(\mathbf{g}_1,\mathbf{g}_2)$ a relative rate. The two new genotypes $\mathbf{g}'_1$ and $\mathbf{g}'_2$ are specified by an indicator variable $\boldsymbol{\xi}$:

$$\mathbf{g}'_1: \quad g_i^{(1)'} = \xi_i g_i^{(1)} + (1-\xi_i) g_i^{(2)} \tag{14}$$

$$\mathbf{g}'_2: \quad g_i^{(2)'} = (1-\xi_i) g_i^{(1)} + \xi_i g_i^{(2)} \tag{15}$$

and this outcome of the recombination happens with probability $C(\boldsymbol{\xi})$. The total rate of the individual event is hence $rQ(\mathbf{g}_1, \mathbf{g}_2)C(\boldsymbol{\xi})$. The change of the distribution over genotypes due to recombination is given by

$$\frac{d}{dt}|_{\text{rec}}P(\mathbf{g}) = r\sum_{\boldsymbol{\xi},\mathbf{g}'} C(\boldsymbol{\xi})\big[Q(\mathbf{g}_1,\mathbf{g}_2)P_2(\mathbf{g}_1,\mathbf{g}_2) - Q(\mathbf{g},\mathbf{g}')P_2(\mathbf{g},\mathbf{g}')\big]. \tag{16}$$

This equation is of a type familiar from non-ideal gas theory: the change in a one-particle distribution (one-genome distribution) depends on the two-particle distributions (two-genome distributions). In practice it is hard to use (16) without a closure, such as assuming that the pair probabilities factorize. Note that the sum on the right-hand side is over one of the parents ($\mathbf{g}'$) and the indicator variable $\boldsymbol{\xi}$ which together give the child $\mathbf{g}$. We assume for simplicity also that Q depends only on the overlap q between the two genotypes $\mathbf{g}$ and $\mathbf{g}'$:

$$q(\mathbf{g},\mathbf{g}') = \frac{1}{L}\sum_{i=1}^{L}\mathbf{1}_{g_i,g'_i}. \tag{17}$$

Recombination as modelled above does not change the overlap. This can be seen as follows: $q(\mathbf{g}_1,\mathbf{g}_2) = 1-\frac{1}{L}\sum_{i=1}^{L}\mathbf{1}_{l_i^{(1)},l_i^{(2)}}$ and $\mathbf{1}_{l_i^{(1)},l_i^{(2)}} = \xi_i(1-\xi_i)\mathbf{1}_{l_i,l_i} + \xi_i^2\mathbf{1}_{l_i,l'_i} + (1-\xi_i)^2\mathbf{1}_{l'_i,l_i} + (1-\xi_i)\xi_i\mathbf{1}_{l'_i,l'_i}$. As the indicator variable takes values zero and one this gives $\mathbf{1}_{l_i^{(1)},l_i^{(2)}} = \mathbf{1}_{l_i,l'_i}$.

Let us now assume that the two-genome distribution factorizes, that the one-genome distribution is of the Potts type and that all quadratic Potts model parameters J_{ij} are small. These assumptions will be seen to be self-consistent when the recombination rate r is high. By a perturbative calculation, which we give in appendix B (essentially the same as found in appendix B of [31]) the right-hand side of (16) simplifies to:

$$\begin{aligned} &r\sum_{\boldsymbol{\xi},\mathbf{g}'} C(\boldsymbol{\xi})P(\mathbf{g}')Q(\mathbf{g},\mathbf{g}')\sum_{i,j,\alpha,\beta}J_{i,j}(\alpha,\beta)\Big(\mathbf{1}_{g_i^{(1)},\alpha}\mathbf{1}_{g_j^{(1)},\beta} \\ &\quad + \mathbf{1}_{g_i^{(2)},\alpha}\mathbf{1}_{g_j^{(2)},\beta} - \mathbf{1}_{g_i,\alpha}\mathbf{1}_{g_j,\beta} - \mathbf{1}_{g'_i,\alpha}\mathbf{1}_{g'_j,\beta}\Big) \\ &= \sum_{i,j,\alpha,\beta} c_{ij}J_{i,j}(\alpha,\beta)\Big(\mathbf{1}_{g_i,\alpha}\mathrm{E}_Q\left[\mathbf{1}_{g'_j,\beta}\right] + \mathrm{E}_Q\left[\mathbf{1}_{g'_i,\alpha}\right]\mathbf{1}_{g_j,\beta} \\ &\quad - \langle Q\rangle\mathbf{1}_{g_i,\alpha}\mathbf{1}_{g_j,\beta} - \mathrm{E}_Q\left[\mathbf{1}_{g'_i,\alpha}\mathbf{1}_{g'_j,\beta}\right]\Big) \end{aligned} \tag{18}$$

where we have used the abbreviations

$$c_{ij} = \sum_{\boldsymbol{\xi}} C(\boldsymbol{\xi})\left(\xi_i(1-\xi_j) + (1-\xi_i)\xi_j\right) \tag{19}$$

$$\langle Q\rangle = \sum_{\mathbf{g}'} Q(\mathbf{g},\mathbf{g}')P(\mathbf{g}') \tag{20}$$

$$\mathrm{E}_Q\left[\mathbf{1}_{g'_i,\alpha}\right] = \sum_{\mathbf{g}'}\mathbf{1}_{g'_i,\alpha}Q(\mathbf{g},\mathbf{g}')P(\mathbf{g}') \tag{21}$$

$$\mathrm{E}_Q\left[\mathbf{1}_{g'_i,\alpha}\mathbf{1}_{g'_j,\beta}\right] = \sum_{\mathbf{g}'}\mathbf{1}_{g'_i,\alpha}\mathbf{1}_{g'_j,\beta}Q(\mathbf{g},\mathbf{g}')P(\mathbf{g}'). \tag{22}$$

The first of these is the probability that two loci are inherited from the same parent and does not (for this model) depend on the genotype $\mathbf{g}$. The last three averages on the other hand depend on $\mathbf{g}$. However, if the function Q is not too sharply focused the dependence can be taken weak. In particular, we assume that $\langle Q\rangle$ is self-averaging, and essentially does not depend on $\mathbf{g}$. In spin glass physics language [44, 45], we hence assume that $\langle Q\rangle$, $\langle Q\mathbf{1}_{l'_i,\alpha}\rangle$ and $\langle Q\mathbf{1}_{l'_i,\alpha}\mathbf{1}_{l'_j,\beta}\rangle$ are self-averaging in the 'paramagnetic' phase where QLE is expected to hold.

5. Evolution equation for log-probability in QLE

In QLE the evolution equation can conveniently be written for the logarithmic probability

$$\frac{d}{dt}\log P(\mathbf{g}) = -\frac{\dot{Z}}{Z} + \sum_{i,\alpha}\dot{h}_i(\alpha)\mathbf{1}_{g_i,\alpha} + \sum_{i,j,\alpha,\beta}\dot{J}_{i,j}(\alpha,\beta)\mathbf{1}_{g_i,\alpha}\mathbf{1}_{g_j,\beta} \tag{23}$$

and the various terms identified. Fitness enters (23) as

$$\frac{d}{dt}|_{\text{fitness}}h_i(\alpha) = F_i(\alpha) \qquad \frac{d}{dt}|_{\text{fitness}}J_{i,j}(\alpha,\beta) = F_{i,j}(\alpha,\beta) \tag{24}$$

and if there were higher-order terms in fitness (more than pair-wise dependencies) they would enter higher than quadratic terms in the QLE distribution in the same way. Ignoring mutations and genetic drift we have for the pair-wise dependencies

$$\dot{J}_{i,j}(\alpha,\beta) = F_{i,j}(\alpha,\beta) - r\langle Q\rangle c_{ij}J_{i,j}(\alpha,\beta) \tag{25}$$

where the contribution from recombination can be read off from (18), for more detail see appendix B. Equation (25) is a relaxation equation which pushes the Potts model parameter $J_{i,j}(\alpha,\beta)$ to be the ratio of two rates, (3) above. When the data is from one population in a stationary state the average relative rate $\langle Q\rangle$ can be subsumed in the overall rate r. When genotypes are Boolean vectors this gives the same result as equation (25) in [31].

As observed above recombination does not change single-locus frequencies, and without mutations the $f_i(\alpha)$ will drift towards fixation (taking values 0 or 1). Once the population has reached fixation at locus i there can no longer be any non-zero correlation of Potts parameter involving i, and in such a setting QLE is therefore only a long-lived quasi-stationary state (for the correlations, and for the $J_{i,j}(\alpha,\beta)$'s). Note that by (18) recombination terms enter in the evolution equations of $h_i(\alpha)$, in combination with the quantities $J_{i,j}(\alpha,\beta)$. This is no contradiction, because when correlations are non-zero there is not a one-to-one relation between single-locus frequencies ($f_i(\alpha)$) and Potts model magnetization parameters ($h_i(\alpha)$); recombination can influence the latter but not the former.

The break-down of the relaxational equation (25) when the single-locus frequencies go to fixation can be understood as follows. In such a setting the $J_{i,j}(\alpha,\beta)$'s would first remain small while the $h_i(\alpha)$'s would tend

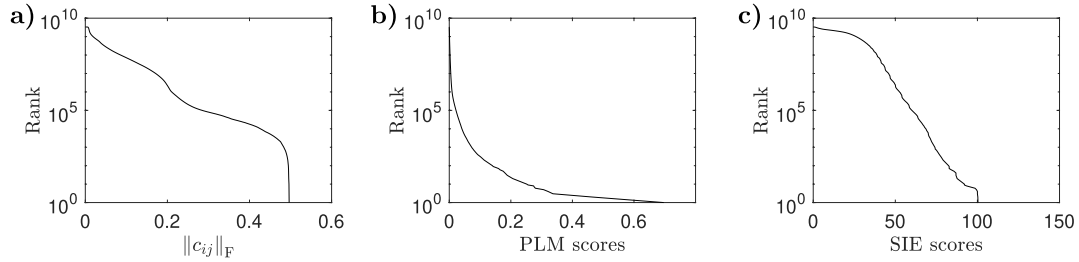

Figure 1. Cumulative distributions of (a) correlations; (b) pseudo-likelihood maximization (PLM); (c) small-interaction expansion (SIE); semi-logarithmic scale visualizing the distributions of of approximately 10^{10} elements. The scalar value associated to each pair of loci i and j is the Frobenius norm of the 3×3-correlation matrix (case a) or the Frobenius norm of the inferred Potts model 3×3-matrix element (cases b and c).

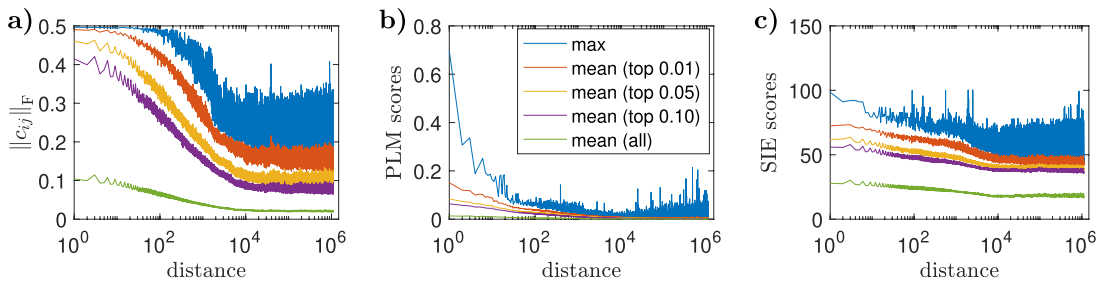

Figure 2. Distributions as functions of genomic distance: (a) correlations; (b) pseudo-likelihood maximization (PLM). (c) Small-interaction expansion (SIE). Same data and same norms as in figure 1. Figures show averages in windows of genomic distance. Blue: maximum, red: top-1%, yellow: top-5%, violet: top-10%, green: mean, all in window. The curves in (b) show a sharp initial decrease with genomic distance, which generally much lower values beyond genomic distance 10^3 where recombination can be expected to act effectively.

to $\pm\infty$. When the $h_i(\alpha)$'s become large enough that the minor alleles in a finite-N population are likely to be present only in a few copies, a few random events can remove all of the remaining ones at once, which sets the correlation and the $J_{i,j}(\alpha, \beta)$ to zero in one go.

6. DCA for whole-genome sequencing data

A set of whole-genome sequences of the human pathogen *S. pneumoniae* obtained in the Maela collaboration (see appendix C) can be represented as about 3000 genotypes of about 100 000 loci each. Correlations and Potts model terms obtained from this data have qualitatively different distributions, as shown in figures 1 and 2.

The number of correlations larger than a cut-off c grows quickly when c decreases below its maximum value, while the cumulative distribution of inferred Potts model couplings has a much more pronounced tail. This implies that the set of largest DCA couplings is better separated than the largest correlations from the unavoidable background due to under-sampling [48, 49]. Correlations also generally have a more uniform distribution across genomic distance, while the representation as a Potts model is sparser (figures 2(a) and (b)). The first-order perturbative version of DCA employed in [30, 31], which is here called SIE and briefly reviewed in appendix A, gives in both instances results closer to correlations (see figures 1(c) and 2(c)).

Figure 3 shows pair-wise density-plots of correlations and DCA terms obtained by PLM and SIE. Because there are too many points, we divide the area inside each frame into 300×300 cells and the colour indicates $\log_{10}(\text{count})$ inside each cell. In all three cases the scatter-plots are 'clouds of points', indicating that DCA and correlations measure different properties of the data. Figure 3(a) shows a weak trend, such that larger PLM scores are associated to larger correlations. Such trends are absent in figures 3(b) and (c), and SIE values are also numerically large. In fact, the correlation matrix is under-sampled, hence smaller correlations reflect sampling noise and this is also an issue for SIE, as well as the sensitive dependence on almost fixed alleles in this procedure. PLM scores and correlations were compared graphically for this data in [24], with a cut-off excluding short-range interactions.

7. The QLE phase is obscured by genetic drift

In a finite population statistical genetics as described above only holds on the average; when following one population in time fluctuations of order $N^{-\frac{1}{2}}$ appear for observables such as single-locus frequencies and pair-wise loci–loci correlations. Figure 4 reports simulations using the `FFPopSim` software showing that these fluctuations can in practice be quite large, even for populations that are not small.

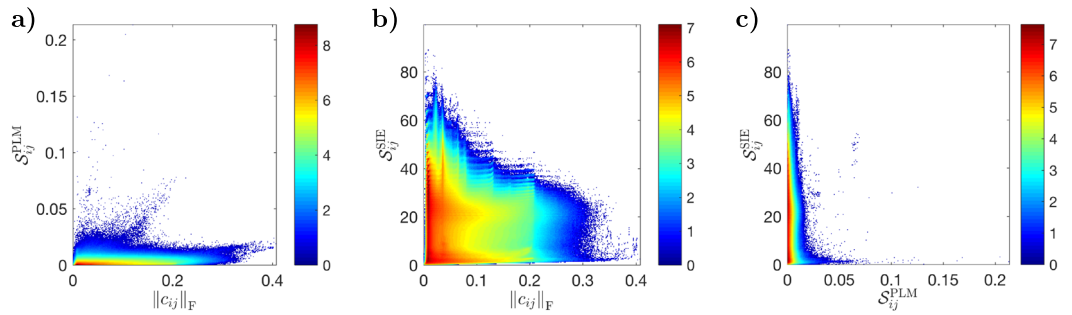

Figure 3. Pair-wise density plot of coupling strength quantified by correlations, PLM scores and SIE scores: (a) PLM scores versus correlations; (b) SIE scores versus correlations; (c) SIE scores versus PLM scores. The area inside axes is divided into 300 × 300 cells and the colour indicates $\log_{10}(\mathrm{count})$ inside each cell. Same data and same norms as in figures 1 and 2. The numerical scale in each direction depends on the details of the norms and inference procedure, e.g. PLM scores depend on ℓ_2 regularization parameters. Correlations and PLM scores are numerically similar while SIE scores are not, as discussed in text.

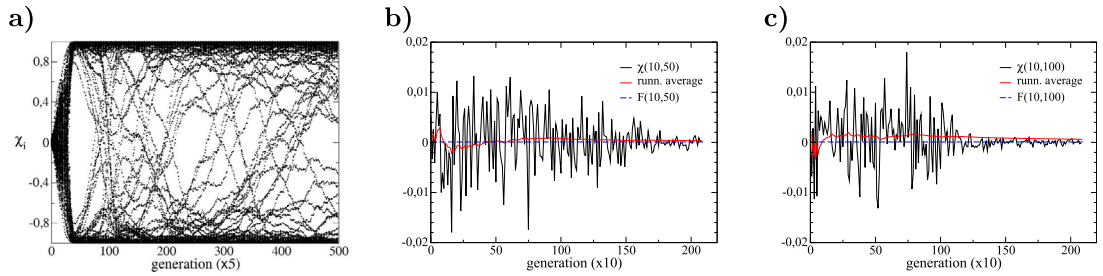

Figure 4. Temporal behavior of (a) all magnetizations defined as $\chi_i = f_i(2) - f_i(1)$ and ((b) and (c)) two selected correlations defined as $\chi_{ij} = f_{ij}(1,1) - f_{ij}(1,2) - f_{ij}(2,1) + f_{ij}(2,2)$ in a simulation of a population of Ising genomes (two alleles per locus). Number of loci L of the genotypes is 256, number of genotypes N in the population is 50 000, other simulation parameters as reported in table F1. Data is taken every five generations, total simulation time is 2500 generations.

According to the theory developed in [31] (appendix C) dynamics of correlations is relaxational and the curves of correlations versus time hence should fluctuate around an equilibrium value, which is the one given in (3). The fluctuations in figure 4 are however large compared to the pair-wise fitness values, and DCA inference from instantaneous values of the ensemble correlations cannot be expected to be good predictors for pairwise fitness. The dynamics of frequencies is not relaxational, and one may hence observe large changes where the population at one locus changes from one allele to another.

8. Discussion

The main question addressed in this work is if and when DCA can be expected to work for genome-scale epistasis analysis. We have given an answer in the context of statistical genetics: for a population evolving under recombination, mutations and fitness variations this is so when recombination is sufficiently fast. The joint distribution of the population over genotypes then approximately factorizes into a product of identical Potts distributions (2). Treating a set of genomes as independent samples from such a distribution allows to infer fitness parameters (F_{ij}) from Potts model parameters (J_{ij}) by inverting (3), and this is essentially what using DCA on genome-scale pneumococcal data means [21, 23, 24, 50]. We now discuss limits to the analysis, and further directions.

A first limitation is that straight-forward application of DCA cannot be expected to yield meaningful results when recombination is weak. One example of such an effect was already given in [21] where also data from *Streptococcus pyogenes* was presented (figure 6 of [21]). Another example was recently given on *Vibrio parahaemolyticus*, a human gastrointestinal pathogen in panmixia, i.e. where all strains are able to recombine, but having a very low overall recombination rate [51]. In contrast, a recent paper reports good results for DCA on *Neisseria gonorrhoeae* [17], a bacterium which also has a high rate of recombination. Similarly, results reported on the one-gene level in HIV [18–20, 52–54] are also explainable by the considerable rate of recombination in the virus when a host is infected by more than one strain [55, 56]. The problem of inferring fitness from data on populations that are in the CC phase or that have only a few large epistatic fitness parameters for which the ratios $\frac{F_{ij}}{rc_{ij}}$ are not small appears to be both conceptually and practically important. We hope to be able to return to such questions in future work.

A second limitation concerns finite populations, particularly simulated data, where the population has to be of moderate size. According to the theory developed for Ising genomes in [31] (appendix C) and qualitatively confirmed above in figure 4, frequencies and correlations follow stochastic differential equations with noise strength scaling as $N^{-\frac{1}{2}}$. In principle Potts model parameters ($h_i(\alpha)$ and $J_{ij}(\alpha,\beta)$) for categorical data also follow stochastic differential equations, but of a more complicated form due to the inverse Ising/Potts relations. Applying the DCA procedure to finite-N data thus requires parameter inference from a high-dimensional stochastic time series with a complicated deterministic part. This may not be an easy task.

A third limitation is the neglect of spatial and environmental separation. Bacteria such as the human pathogen *Helicobacter pylori* readily recombine if they meet, but can only do so when their human host populations overlap [57]. Allele frequencies may be different for different bacterial populations, reflecting differences in the host populations and environments. If data from the different populations is pooled, this will be a confounding factor for some flavors of DCA e.g. for PLM and SIE, these types of issues appear to merit further study.

Acknowledgments

EA thanks Prof Boris Shraiman for enlightening discussions. This research was supported by National Science Foundation of China (grant numbers 11647601 and 11421063) and by ESPCI Chaire Joliot 2018 (EA). The numerical computations were partly carried out at the HPC cluster of ITP-CAS.

Appendix A. The small-interaction expansion (SIE) for categorical data

The small-interaction expansion (SIE) is introduced for Boolean data in [30, 31]. Here we give a derivation of SIE for categorical data. We need the solution of the matrix equation $\mathbf{u}_\alpha = \sum_\beta M_{ii}(\alpha,\beta)\mathbf{v}_\beta$ in the space of vectors orthogonal to $(1,1,\ldots,1)$, where the one-locus allele correlation matrix $M_{ii}(\alpha,\beta)$ is defined in (7). That is given by

$$v_\alpha = \frac{u_\alpha}{f_i(\alpha)} - \frac{1}{A_i}\sum_{\beta=1}^{A_i}\frac{u_\beta}{f_i(\beta)}. \tag{A.1}$$

Consider now the Potts model when all the interaction parameters $J_{ij}(\alpha,\beta)$ are small. One frequency can be estimated to zeroth and first order as

$$f_i(\alpha) = \frac{e^{h_i(\alpha)}}{N_i} + \sum_{j,\beta} J_{i,j}(\alpha,\beta)\frac{e^{h_i(\alpha)}}{N_i}\frac{e^{h_j(\beta)}}{N_j} - \sum_{j,\beta,\gamma} J_{i,j}(\gamma,\beta)\frac{e^{h_i(\alpha)}}{N_i}\frac{e^{h_i(\gamma)}}{N_i}\frac{e^{h_j(\beta)}}{N_j} \tag{A.2}$$

where we have used the abbreviation $N_i = \sum_\alpha e^{h_i(\alpha)}$. The fluctuation–dissipation relations for the Potts model read

$$M_{ij}(\alpha,\beta) = \frac{\partial f_i(\alpha)}{\partial h_j(\beta)} \tag{A.3}$$

and therefore, comparing (7),

$$M_{ij}(\alpha,\beta) = \sum_{\gamma,\delta} J_{i,j}(\gamma,\delta)\, M_{ii}(\alpha,\gamma)\, M_{jj}(\beta,\delta). \tag{A.4}$$

Since the Potts parameters are in the Ising gauge [2, 47] the matrix multiplications in (A.4) can be inverted using (A.1):

$$J_{i,j}(\alpha,\beta) = \frac{M_{ij}(\alpha,\beta)}{f_i(\alpha)f_j(\beta)} - \frac{1}{f_j(\beta)}\sum_\gamma \frac{M_{ij}(\gamma,\beta)}{f_i(\gamma)} - \frac{1}{f_i(\alpha)}\sum_\delta \frac{M_{ij}(\alpha,\delta)}{f_j(\delta)} + \sum_{\gamma,\delta}\frac{M_{ij}(\gamma,\delta)}{f_i(\gamma)f_j(\delta)}. \tag{A.5}$$

This can be interpreted as an inference algorithm where the interaction parameters $J_{ij}(\alpha,\beta)$ are determined from the single-locus frequencies $f_i(\alpha)$ and pairwise locus–locus correlations $M_{ij}(\alpha,\beta)$. Since it is a first-order perturbative solution to naive mean-field inference, SIE can be expected to be a comparatively weak inference procedure. Figure A1 confirms that this is the case for the Sherrington–Kirkpatrick spin glass model. An implementation of SIE for categorical data can be found at www.github.com/gaochenyi/DCA-QLE.

Appendix B. Perturbative calculation of recombination in QLE phase

This appendix contains a detailed derivation of (18) in the main text. The derivation is analogous to the derivation of equation (B4 *b*) in [31], with modifications due to categorical data and our model for the recombination process.

The starting point is the right hand side of (16) which we repeat here up to the overall factor r:

$$\text{Expr.} = \sum_{\boldsymbol{\xi},\mathbf{g}'} C(\boldsymbol{\xi})\big[Q(\mathbf{g}_1,\mathbf{g}_2)P_2(\mathbf{g}_1,\mathbf{g}_2) - Q(\mathbf{g},\mathbf{g}')P_2(\mathbf{g},\mathbf{g}')\big]. \tag{B.1}$$

It is understood in (B.1) $\mathbf{g}_1$ and $\mathbf{g}_2$ are two parent genomes that give rise to two child genomes $\mathbf{g}$ and $\mathbf{g}'$ related to the parent genomes by the indicator variable $\boldsymbol{\xi}$.

Let us now assume that *(i)* the two-genome distribution functions factorize and *(ii)* the one-genome distribution functions are of the Potts type. Then

$$P_2(\mathbf{g}_1,\mathbf{g}_2) = \frac{1}{Z^2}\exp\left(\sum_i\left(h_i(g_i^{(1)}) + h_i(g_i^{(2)})\right)\right) \cdot \exp\left(\sum_{ij}\left(J_{ij}(g_i^{(1)},g_j^{(1)}) + J_{ij}(g_i^{(2)},g_j^{(2)})\right)\right).$$

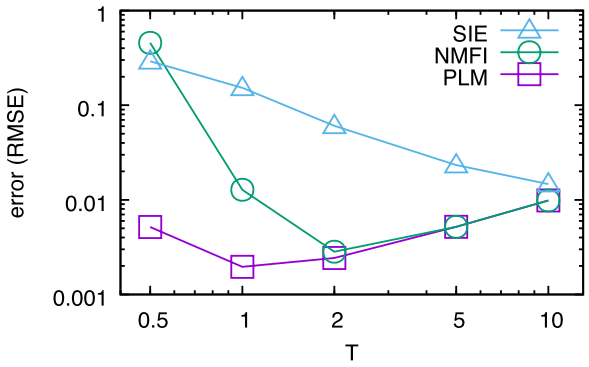

Figure A1. Root mean square error of three inference algorithms on Sherrington–Kirkpatrick (SK) spin glass. Abscissa (x-axis): temperature. Ordinate (y-axis): pseudo-likelihood maximization (PLM), naive mean-field inversion (NMFI) and small-interaction expansion (SIE). The SK model is a widely studied test case in the DCA literature; performance of other inference algorithms can be found in *see* [1], and references cited therein.

According to the indicator function $\boldsymbol{\xi}$ the value of g_i is $g_i^{(1)}$ or $g_i^{(2)}$, the other value taken by g_i'. Hence

$$h_i(g_i^{(1)}) + h_i(g_i^{(2)}) = h_i(g_i) + h_i(g_i') \tag{B.2}$$

the same effect that recombination does not change single-locus frequencies. We can therefore write

$$P_2(\mathbf{g}_1, \mathbf{g}_2) = P_2(\mathbf{g}, \mathbf{g}') \exp\left(\sum_{ij}\left(J_{ij}(g_i^{(1)}, g_j^{(1)}) + J_{ij}(g_i^{(2)}, g_j^{(2)})\right)\right)$$
$$\cdot \exp\left(\sum_{ij}\left(-J_{ij}(g_i, g_j) - J_{ij}(g_i', g_j')\right)\right). \tag{B.3}$$

If we now assume *(iii)* that all the quadratic Potts terms are small we can expand the exponential in (B.3) to zeroth and first order in the quantities J_{ij}.

Using that Q by assumption only depends on the overlap and that overlap is preserved by recombination the zeroth order contribution to (B.1) is

$$\text{Expr.} = P(\mathbf{g}) \sum_{\boldsymbol{\xi}, \mathbf{g}'} C(\boldsymbol{\xi}) \left[Q(\mathbf{g}, \mathbf{g}')P(\mathbf{g}') - Q(\mathbf{g}, \mathbf{g}')P(\mathbf{g}')\right] \tag{B.4}$$

which vanishes. The first order contributions on the other hand give (18) in the main text.

Appendix C. *S. pneumoniae* sequence data

Whole-genome sequences of carriage isolates from two birth cohorts of infants and their mothers in the Maela refugee camp (Thailand) [58, 59] were reported in [41]. This data was filtered for positions (loci) that carry at most two alleles and a moderate amount of gaps, as described previously [21, 24]. This procedure results in 3145 genotypes each containing 81 506 loci, where the alleles at each locus can take three values (major, minor, gap). The original MSA data can be found in [21], while the filtered MSA can be retrieved by the pipeline function in www.github.com/gaochenyi/DCA-QLE.

Appendix D. Correlation matrix computations

Correlation matrices were computed using the MATLAB implementation available at [60] (www.github.com/gaochenyi/CC-PLM). On the Maela data set ($L = 10^5$) the compute time was approximately 30 core-hours using a 56-core server with four Intel Xeon E7-4850 v3 processors. The run-time memory used is about 70 GB storing all correlations in memory.

Appendix E. Direct coupling analyses

Potts model parameters were inferred by the asymmetric ℓ_2-regularized pseudo-likelihood maximization method [47] using the software PLM at [60] (www.github.com/gaochenyi/CC-PLM). On the same data set and in the same compute environment as above, the total compute time was about 20 000 core-hours. The implementation of naive mean-field inference (NMFI) for Boolean data used in the paper can be found at www.github.com/gaochenyi/DCA-QLE.

Appendix F. Simulations of Wright–Fisher model with pairwise fitness function and recombination

Simulations of the Wright–Fisher model with recombination were done with the `FFPopSim` simulation package [61] with parameter settings as given in table F1.

In the simulations reported in figure 4 single-locus contributions to fitness (parameters F_i) are zero, while pair-wise loci–loci contributions (parameters F_{ij}) are random and small ($\pm 7.8 \times 10^{-5}$).

Table F1. Parameter settings for the simulations reported as figure 4 in main text.

Parameter	Setting
Number of loci (L)	256
Circular	Yes
Number of traits	1
Population size	126 641
Carrying capacity (N)	50 000
Generation	0
Outcrossing rate (r)	1.0
Crossover rate (ρ)	0.05
Recombination model	CROSSOVERS
Mutation rate (μ)	0.01
Participation ratio Y	0.0022
Number of non-empty clones (N)	500

ORCID iDs

Erik Aurell https://orcid.org/0000-0003-4906-3603

References

[1] Nguyen H C, Zecchina R and Berg J 2017 Inverse statistical problems: from the inverse Ising problem to data science *Adv. Phys.* **66** 197

[2] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 Identification of direct residue contacts in protein–protein interaction by message passing *Proc. Natl Acad. Sci. USA* **106** 67

[3] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families *Proc. Natl Acad. Sci. USA* **108** E1293

[4] Stein R R, Marks D S and Sander C 2015 Inferring pairwise interactions from biological data using maximum-entropy probability models *PLoS Comput. Biol.* **11** e1004182

[5] Michel M, Skwark M J, Menéndez Hurtado D, Ekeberg M and Elofsson A 2017 Predicting accurate contacts in thousands of Pfam domain families using PconsC3 *Bioinformatics* **33** 2859

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

[7] Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos G A, Kim D E, Kamisetty H, Kyrpides N C and Baker D 2017 Protein structure determination using metagenome sequence data *Science* **355** 294

[8] Michel M, Menéndez Hurtado D, Uziela K and Elofsson A 2017 Large-scale structure prediction by improved contact predictions and model quality assessment *Bioinformatics* **33** i23

[9] Ovchinnikov S, Park H, Kim D E, DiMaio F and Baker D 2018 Protein structure prediction using Rosetta in CASP12 *Proteins* **86** 113

[10] De Leonardis E, Lutz B, Ratz S, Simona C, Monasson R, Weigt M and Schug A 2015 Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction *Nucleic Acids Res.* **43** 10444

[11] Gueudré T, Baldassi C, Zamparo M, Weigt M and Pagnani A 2016 Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis *Proc. Natl Acad. Sci. USA* **113** 12186

[12] Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H and Weigt M 2017 Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis *Proc. Natl Acad. Sci. USA* **114** E2662

[13] Weinreb C, Riesselman A J, Ingraham J B, Gross T, Sander C and Marks D S 2016 3D RNA and functional interactions from evolutionary couplings *Cell* **165** 963

[14] Figliuzzi M, Jacquier H, Schug A, Tenaillon O and Weigt M 2016 Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1 *Mol. Biol. Evol.* **33** 268

[15] Hopf T A, Ingraham J B, Poelwijk F J, Scharfe C P I, Springer M, Sander C and Marks D S 2017 Mutation effects predicted from sequence co-variation *Nat. Biotechnol.* **35** 128

[16] Couce A, Caudwell L V, Feinauer C, Hindré T, Feugeas J-P, Weigt M, Lenski R E, Schneider D and Tenaillon O 2017 Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria *Proc. Natl Acad. Sci. USA* **114** E9026

[17] Schubert B, Maddamsetti R, Nyman J, Farhat M R and Marks D S 2018 Genome-wide discovery of epistatic loci affecting antibiotic resistance using evolutionary couplings *Nat. Microbiol.* **4** 328–38

[18] Ferguson A L, Mann J K, Omarjee S, Ndung'u T, Walker B D and Chakraborty A K 2013 Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design *Immunity* **38** 606

[19] Shekhar K, Ruberman C F, Ferguson A L, Barton J P, Kardar M and Chakraborty A K 2013 Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes *Phys. Rev.* E **88** 062705

[20] Louie R H Y, Kaczorowski K J, Barton J P, Chakraborty A K and McKay M R 2018 Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies *Proc. Natl Acad. Sci. USA* **115** E564

[21] Skwark M J *et al* 2017 Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis *PLoS Genet.* **13** e1006508

[22] Hakenbeck R, Brückner R, Denapaite D and Maurer P 2012 Molecular mechanisms of β-lactam resistance in *Streptococcus pneumoniae Future Microbiol.* **7** 395

[23] Puranen S, Pesonen M, Pensar J, Xu Y Y, Lees J A, Bentley S D, Croucher N J and Corander J 2018 SuperDCA for genome-wide epistasis analysis *Microbial Genomics* **4**

[24] Gao C-Y, Zhou H-J and Aurell E 2018 Correlation-compressed direct-coupling analysis *Phys. Rev.* E **98** 032407

[25] Aurell E 2016 The maximum entropy fallacy redux? *PLoS Comput. Biol.* **12** 1

[26] van Nimwegen E 2016 Inferring contacting residues within and between proteins: what do the probabilities mean? *PLoS Comput. Biol.* **12** e1004726

[27] Kimura M 1956 A model of a genetic system which leads to closer linkage by natural selection *Evolution* **10** 278

[28] Kimura M 1964 Diffusion models in population genetics *J. Appl. Probab.* **1** 177

[29] Kimura M 1965 Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection *Genetics* **52** 875

[30] Neher R A and Shraiman B I 2009 Competition between recombination and epistasis can cause a transition from allele to genotype selection *Proc. Natl Acad. Sci. USA* **106** 6866

[31] Neher R A and Shraiman B I 2011 Statistical genetics and evolution of quantitative traits *Rev. Mod. Phys.* **83** 1283

[32] Fisher R 1922 On the dominance ratio *Proc. R. Soc. Edinburgh* **42** 321

[33] Fisher R 1930 *The Genetical Theory of Natural Selection* (Oxford: Clarendon)

[34] Kolmogorov A N 1935 Deviations from hardy's formulas under partial isolation *Dokl. Akad. Nauk SSSR* **3** 129

[35] Peliti L 1997 Introduction to the statistical theory of Darwinian evolution *Lectures at the Summer College on Frustrated System (Trieste, August 1997)* (arXiv:cond-mat/9712027)

[36] Blythe R A and McKane A J 2007 Stochastic models of evolution in genetics, ecology and linguistics *J. Stat. Mech.* P07018

[37] Mustonen V and Lässig M 2010 Fitness flux and ubiquity of adaptive evolution *Proc. Natl Acad. Sci. USA* **107** 4248

[38] Roach J C *et al* 2010 Analysis of genetic inheritance in a family quartet by whole-genome sequencing *Science* **328** 636

[39] Chaguza C et al 2016 Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule *mBio* **7** e01053
[40] Yahara K, Didelot X, Ansari M A, Sheppard S K and Falush D 2014 Efficient inference of recombination hot regions in bacterial genomes *Mol. Biol. Evol.* **31** 1593
[41] Chewapreecha C et al 2014 Dense genomic sampling identifies highways of pneumococcal recombination *Nat. Genet.* **46** 305
[42] Neher R A, Vucelja M, Mézard M and Shraiman B I 2013 Emergence of clones in sexual populations *J. Stat. Mech.* P01008
[43] Dikmen O 2015 Learning mixtures of Ising models using pseudolikelihood (arXiv:1506.02510)
[44] Méezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and beyond: an Introduction to the Replica Method and its Applications* (Singapore: World Scientific)
[45] Méezard M and Montanari A 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)
[46] Lewontin R C 1964 The interaction of selection and linkage. I. General considerations; heterotic models *Genetics* **49** 49
[47] Ekeberg M, Hartonen T and Aurell E 2014 Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences *J. Comput. Phys.* **276** 341
[48] Xu Y, Aurell E, Corander J and Kabashima Y 2017 Statistical properties of interaction parameter estimates in direct coupling analysis (arXiv:1704.01459 [physics.data-an])
[49] Xu Y, Puranen S, Corander J and Kabashima Y 2018 Inverse finite-size scaling for high-dimensional significance analysis *Phys. Rev.* E **97** 062112
[50] Zhang J 2017 Epistasis analysis goes genome-wide *PLoS Genet.* **13** e1006558
[51] Cui Y, Yang C, Qiu H, Wang H, Yang R and Falush D 2018 The landscape of coadaptation in *Vibrio parahaemolyticus bioRxiv* 373936
[52] Barton J P, Goonetilleke N, Butler T C, Walker B D, McMichael A J and Chakraborty A K 2016 Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable *Nat. Commun.* **7** 11660
[53] Butler T C, Barton J P, Kardar M and Chakraborty A K 2016 Identification of drug resistance mutations in HIV from constraints on natural evolution *Phys. Rev.* E **93** 022412
[54] Chakraborty A K and Barton J P 2017 Rational design of vaccine targets and strategies for HIV: a crossroad of statistical physics, biology, and medicine *Rep. Prog. Phys.* **80** 032601
[55] Burke D S 1997 Recombination in HIV: an important viral evolutionary strategy *Emerg. Infectious Dis.* **3** 253–9
[56] Neher R A and Leitner T 2010 Recombination rate and selection strength in HIV intra-patient evolution *PLoS Comput. Biol.* **6** e1000660
[57] Thorell K et al 2017 Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas *PLoS Genet.* **13** 1–21
[58] Turner P, Turner C, Jankhot A, Helen N, Lee S J, Day N P, White N J, Nosten F and Goldblatt D 2012 A longitudinal study of *Streptococcus pneumoniae* carriage in a cohort of infants and their mothers on the Thailand–Myanmar border *PLoS One* **7** e38271
[59] Turner C, Turner P, Carrara V, Burgoine K, Htoo S T L, Watthanaworawit W, Day N P, White N J, Goldblatt D and Nosten F 2013 High rates of pneumonia in children under two years of age in a south east asian refugee population *PLoS One* **8** e54026
[60] Gao C-Y 2018 Github www.gaochenyi/CC-PLM
[61] Neher R and Zanini F 2012 FFPopSim http://code.google.com/p/ffpopsim/