

# Predicting evolution

Michael Lässig<sup>1\*</sup>, Ville Mustonen<sup>2\*†</sup> and Aleksandra M. Walczak<sup>3\*</sup>

**The face of evolutionary biology is changing: from reconstructing and analysing the past to predicting future evolutionary processes. Recent developments include prediction of reproducible patterns in parallel evolution experiments, forecasting the future of individual populations using data from their past, and controlled manipulation of evolutionary dynamics. Here we undertake a synthesis of central concepts for evolutionary predictions, based on examples of microbial and viral systems, cancer cell populations, and immune receptor repertoires. These systems have strikingly similar evolutionary dynamics driven by the competition of clades within a population. These dynamics are the basis for models that predict the evolution of clade frequencies, as well as broad genetic and phenotypic changes. Moreover, there are strong links between prediction and control, which are important for interventions such as vaccine or therapy design. All of these are key elements of what may become a predictive theory of evolution.**

Chance and necessity of evolution are a classic topic in biology<sup>1–4</sup>. Because evolution is shaped by multiple stochastic forces of reproduction and environment, many have taken a sceptical view on its repeatability, let alone predictability. With only sparse and predominantly macro-evolutionary data, this question has, for a long time, remained difficult to decide by direct comparison of experiment and theory. The situation is now changing; massively parallel evolution experiments, high-throughput sequencing and phenotypic assays, and progress in modelling complex dynamical processes provide an unprecedented amount of evolutionary information. The new data and methods paint a more upbeat picture of predictability in evolution, albeit on shorter time scales. They reveal that evolutionary processes show repeatable features: different pathogen populations evolve similar resistance to a given antibiotic, immune systems of different hosts evolve similar receptors against the same pathogen, and cancers are marked across patients by mutations in specific oncogenes<sup>5–11</sup>. Building on these regularities, a number of recent studies have come up with actual predictions of future evolution in specific systems<sup>12–16</sup>.

What is predictable in evolution, what may become predictable in the near future, and what will remain unpredictable? These are the central questions of this Perspective. Here we use the term prediction in a specific sense: a testable hypothesis about an evolutionary process that extends into the future. This distinguishes evolutionary predictions from the broader usage of the term prediction (of a model, that can be tested by experiment) and excludes processes with solely metabolic or ecological dynamics. Building on recent progress in microbial and viral evolution, cancer evolution, and the somatic evolution of immune systems, we develop unifying concepts for predictive analysis and identify avenues for future research.

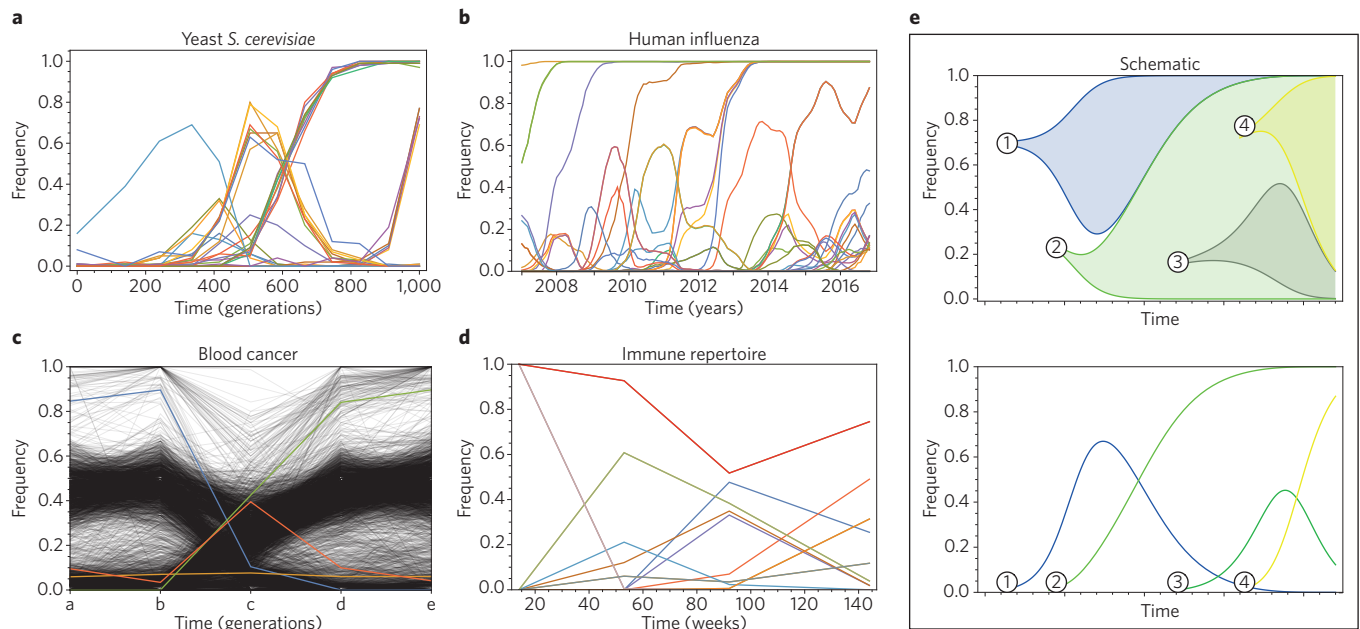
## What makes evolution predictable?

A look at evolutionary processes on the molecular scale seems to support scepticism on predictability. Molecular evolution is driven by mutations that arise randomly in an individual's genome and act on complicated, in part unknown cellular machinery. The fate of mutations in an evolving population appears similarly complicated. In Fig. 1a–d, we plot the frequency paths of genetic mutations in

systems representative for this article, which include a laboratory population of yeast cells, the human influenza virus A/H3N2, and populations of cancer and immune B cells in a human individual. These systems are examples of Darwinian evolution: genetic variation is continuously produced by mutations and is acted upon by selection. Part of the positively selected changes expands in the entire population and generates increasing divergence from its initial state. A closer look reveals the complexity of the evolutionary dynamics. All of the populations have multiple coexisting clades (that is, groups of genetically related individuals); beneficial mutations in disjoint clades compete for fixation, while mutations in nested clades reinforce one another (Fig. 1e). This evolutionary mode, which is commonly called clonal interference, arises in large asexual populations subject to strong selection<sup>17</sup>. Here we use the term clades (instead of clones) to highlight that successful clades acquire new genetic diversity on their way to fixation (Fig. 1e). Clonal interference has been observed in laboratory evolution of microbial and viral populations<sup>18,19</sup> and probably governs all of the systems shown in Fig. 1<sup>20–25</sup>. Two of its characteristics are relevant for predictions. First, new mutants are produced at a high rate, which reduces stochastic waiting times for fitter genetic variants. Second, the observed rise and decline of clade frequencies is driven by selection, not by genetic drift or environmental noise. Fitness models, that is, models that estimate selection on clades from past evolutionary data, can rationalize these dynamics and, at least in principle, predict future changes. Thus, the very factors that generate the complexity of the evolutionary process enhance its repeatability and hold the key to predictive analysis.

If selection is to generate predictability, it must prune a highly complex space of evolutionary possibilities to essentially a single likely alternative. Research in recent years has revealed that at different levels of biological organization, the degree to which this takes place varies greatly. In parallel-evolving laboratory populations, the vast majority of single-nucleotide and amino acid changes occur in just a single population<sup>26–29</sup>; that is, different populations follow divergent paths in sequence space (Fig. 2a). Somatic evolution in multicellular organisms, which is one of nature's massively parallel evolution experiments, shows a similar picture of genetic heterogeneity. Cancer genotypes, even for tumours of the same type, have

<sup>1</sup>Institute of Theoretical Physics, University of Cologne, 50937 Cologne, Germany. <sup>2</sup>Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK. <sup>3</sup>Laboratoire de Physique Théorique CNRS, Ecole Normale Supérieure, 75005 Paris, France. <sup>†</sup>Present address: Department of Biosciences, University of Helsinki, PO Box 65, 00014, Finland. \*e-mail: mlaessig@uni-koeln.de; vm5@sanger.ac.uk; awalczak@lpt.ens.fr



**Figure 1 | Clonal interference is a common mode of evolution.** Time-dependent frequencies of genetic variants are shown for four systems. **a**, Yeast cells under clonal evolution for 1,000 generations<sup>38</sup>. Several sets of mutations first increase in frequency but are eventually outcompeted by another lineage. **b**, The human influenza lineage A/H3N2 in the period 2007–16<sup>13</sup>. This plot shows strong competition between viral clades and the fixation of several epitope mutations. **c**, Blood cancer cells sampled at five time-points over a period of three years<sup>23</sup>. Clone frequencies (coloured lines) are inferred<sup>141</sup> from frequencies of somatic single nucleotide variants (black lines; these frequencies differ because cancer cells are diploid and contain copy number changes). **d**, Immune B cells evolving the broadly neutralizing HIV antibody CH103 sampled at four time-points over a period of 140 weeks<sup>142</sup>. The B cell clade carrying a positively selected precursor of a broadly neutralizing antibody originates in week 93 (orange line) and rises in frequency together with two parental clades (red lines). All of these populations continuously harbour substantial genetic and fitness differences, which form the basis for evolutionary predictions. **e**, The observed mutation patterns signal a specific mode of Darwinian evolution: beneficial mutations in disjoint clades compete (for example, 1 and 2), beneficial mutations in nested clades reinforce one another (for example, 2 and 3). Top, Muller plot displaying ancestor and frequency (height difference enclosed by the shaded area) for each clade. Bottom, corresponding mutation frequency plot.

largely disjoint spectra of mutations between patients<sup>30</sup>. Already the starting points of cancer evolution are diverse; the somatic evolution even of healthy tissues produces genetic heterogeneity<sup>31</sup>. Another important somatic process is the generation of adaptive immune receptors in vertebrates. Combinatoric assembly of genomic templates, followed by random nucleotide insertions and deletions, produces repertoires that are staggeringly diverse within and largely disjoint between individuals<sup>32–34</sup>. Consequently, different individuals respond to an infection or vaccination by different immune receptor sequences<sup>35</sup>.

The divergence of genome evolution observed in all of these systems is hardly surprising. Even simple units of biological systems have a large number of possible mutational changes that have similar functional effects. For example, loss of genes in regulatory networks is frequently observed in evolution experiments<sup>36</sup>, and a given gene can be silenced by many different sequence mutations. More generally, changes in gene regulation and in cell metabolism have a large mutational target, given the redundancies in regulatory sequence grammar and metabolic pathways. These redundancies imply that ‘microscopic’ genome evolution is not repeatable. But they also hold a positive message for predictability: in order to forecast functional changes in a population, we do not need to know the exact evolutionary path in sequence space.

At a more coarse-grained level, recent evolution experiments do suggest a route towards predictive analysis. Heritable phenotypes, in particular quantitative traits, often evolve in a more repeatable way than genomic sequences. This is a common feature of microbial populations<sup>4,27,37–40</sup>, where repeatable changes occur in regulatory or metabolic pathways. Such changes are promising building blocks for predictive analysis. They can be tracked by observations of genetic

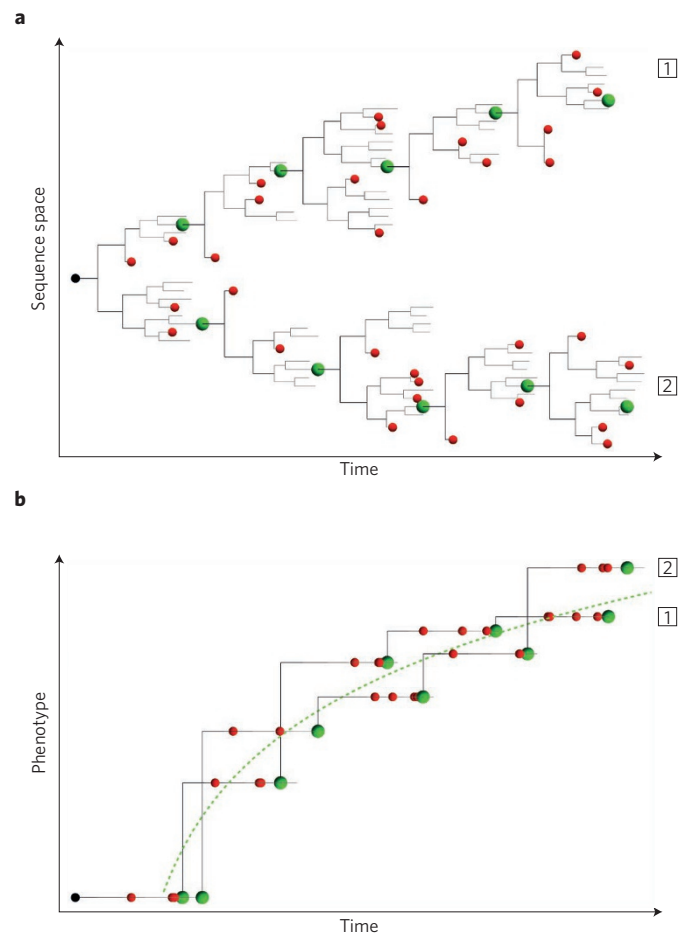
convergence: in parallel-evolving populations, mutations frequently affect the same genes, operons, or larger functional units<sup>27,41,42</sup>. In adaptive processes, high-level organismic traits, in particular fitness itself, can even produce a highly regular pattern of time dependence<sup>20,26</sup> (Fig. 2b). The evolution of cancer and of adaptive immune repertoires follows a similar pattern. Cancer is marked by a series of high-level changes, including resistance to cell death and sustained proliferative signalling<sup>43</sup>. These phenotypic changes have a large sequence target and are highly repeatable, to an extent they can serve as a definition of cancer. More specific phenotypes, such as transcriptional states, appear to be informative of disease progression<sup>44</sup>. At the peak of an acute infection, a substantial fraction of the functional immune receptors responds to specific antigens<sup>22,45</sup>. This repeatable adaptive evolution is mediated by immune receptor–antigen affinity phenotypes that may be identifiable from characteristic sequence motifs<sup>34,46</sup> (similar to the well-known sequence motifs of transcriptional regulation).

The differences in repeatability between genomic and phenotypic evolution reflects the dependence of selection on biological scale. Sequence space contains a staggering number of evolutionary paths. Although negative and positive selection reduces the number of likely paths, sequence evolution remains generically unrepeatable (Fig. 2a). Two main factors generate stochasticity: mutations with similar functional effects have similar fitness effects and similar likelihood; moreover, a fraction of the system’s genomic sites evolves under weak selection altogether. For example, sites that are part of quantitative traits with sequence redundancy evolve near neutrality, even if the trait itself is under substantial stabilizing selection. More generally, clonal interference acts as a selective filter: only strongly adaptive mutations are governed by their own

selection coefficient and can evolve repeatably; moderately selected beneficial and deleterious mutations acquire near-neutral fixation probabilities<sup>47,48</sup> and lose repeatability. This effect has been called emergent neutrality<sup>47,48</sup> and can be understood from Fig. 1e: moderately selected mutations are pushed and pulled by the dynamics of clades in their immediate back- and foreground, which is driven by stronger selection. The theoretical expectation of emergent neutrality is in line with the hitchhiking of deleterious mutations observed in evolution experiments and wild populations<sup>29,49–52</sup>. At the phenotypic level, selection acts in a stronger and more coherent way. First, negative (stabilizing) selection and physiological constraints play an important role in reducing the number of phenotypic evolutionary paths<sup>53</sup>. Stabilizing selection, for example, on protein fold stability or catalytic activity, is a general feature of protein evolution, which has been inferred for several microbial and viral systems<sup>54–62</sup>. More broadly, the collective effect of deleterious changes matters for predictions despite their individual stochasticity: a substantial fitness cost has been observed, for example, in cancer<sup>63,64</sup> and influenza<sup>13,21,65</sup>. Second, positive (directional) selection further prunes the number of evolutionary paths in adaptive processes. Most processes discussed here fall into this class, notably the evolutionary response of microbial, viral, and cancer populations to drugs and of immune repertoires to antigens. In the adaptive evolution of antibiotic resistance<sup>66–69</sup>, repeatability has been observed for some large-effect sequence changes, but more broadly for molecular resistance phenotypes<sup>39,66–72</sup>. Together, an evolutionary process is, in principle, predictable to the extent that (negative and positive) selection canalizes phenotypic evolution towards a single dominant path (Fig. 2b). We will discuss below how this criterion plays out in practice. The joint role of conservation and adaptation in reducing evolutionary complexity extends to macro-evolution; an example is gene regulatory networks that establish new links while maintaining ancestral functions<sup>72,73</sup>.

Phenotypic evolution is marked by correlations that can be harvested for the inference of fitness landscapes and for predictive analysis. One source of such correlations is the nonlinearity of phenotypic fitness landscapes, which implies broad fitness interactions (epistasis) between mutations: deleterious changes increase in cost with increasing distance from a ridge; beneficial changes decrease in return with decreasing distance from a peak<sup>74</sup>. Importantly, these interactions generate evolutionary constraints and increase the predictability of phenotypic processes and outcome. Yeast populations, for example, show a rate of adaptation that is predictable in terms of their initial fitness<sup>38</sup>. Even in macro-evolutionary processes, phenotypic epistasis can generate a predictable order of evolutionary steps, as has been observed in the evolution of complex functions in prokaryotes by lateral gene transfer<sup>75</sup> and of photosynthesis in plants<sup>76</sup>. The emerging picture of smooth phenotype-fitness maps with ‘macroscopic’ epistasis<sup>38</sup> (Fig. 2b) is in sharp contrast to that of rugged fitness landscapes on sequence space. The latter are dominated by ‘microscopic’ epistasis, which decreases the number of accessible evolutionary paths<sup>66,72</sup>, but the local peaks and valleys in a given system can hardly be captured by a predictive model with few parameters<sup>70,77–80</sup>. We conclude that microscopic and macroscopic epistasis can both enhance repeatability but have opposing effects on predictability. This illustrates an important general point: repeatability is a necessary but not a sufficient condition for predictability.

In the densely packed genomes of microbial and viral systems, multiple traits are often encoded in common genetic loci. This property (called pleiotropy) is another source of evolutionary correlations relevant for predictions. Pleiotropy constrains adaptive evolution to characteristic serpentine paths: primary beneficial mutations advance adaptive traits but degrade conserved traits encoded at the same site (because the adaptive allele is, on average, deleterious for other traits). The collateral damage of adaptation is subsequently repaired by compensatory mutations<sup>55–57,59,61,81–87</sup> (Fig. 3a). A similar

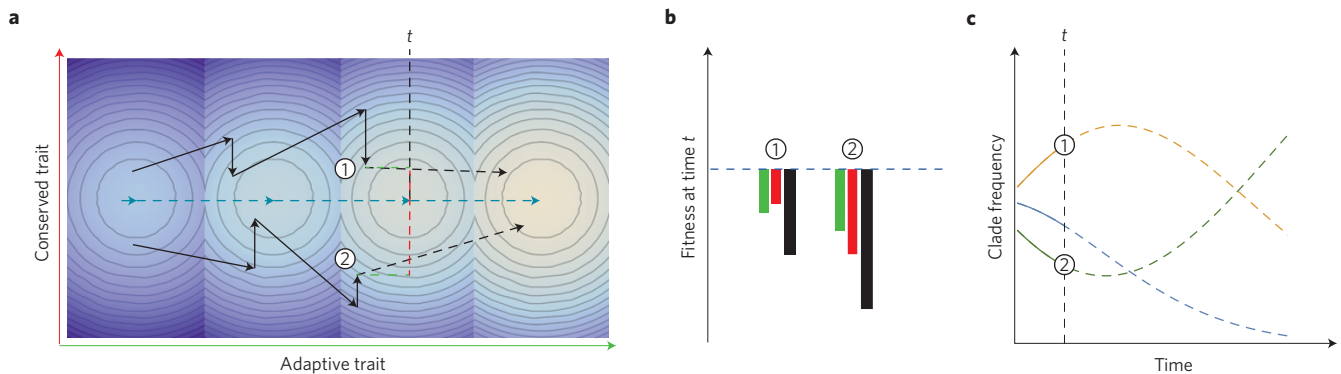


**Figure 2 | Predictability in evolution (schematic).** The figure compares evolutionary paths at different biological scales for an adaptive process in two parallel populations (1 and 2) with the same initial state (black dot). **a**, Evolution in sequence space is a stochastic process with many equiprobable paths. Mutations continuously generate new genetic variants, shown as nodes on a sequence tree. This process is only partially constrained by selection (red dots, mutations under negative selection; green dots, mutations under positive selection). Hence, parallel populations follow different paths on a sequence tree. Detailed genetic evolution is not repeatable and cannot be predicted. **b**, Fitness and other organismic traits of the same process can evolve in a highly regular way. The phenotypic paths shown contain the same mutations (red and green dots) as in **a**; different sequence changes map onto similar phenotypic effects. Negative selection and adaptive pressure together may canalize evolution towards a single phenotypic path (dashed green line). Parallel populations evolve close to this path. Such processes are repeatable and, in principle, predictable. Evolutionary control produces adaptive pressure and can enhance predictability by specific protocols of experiment or treatment.

dynamics arises if conserved traits are affected by strong hitchhiking of their genetic loci. Compensatory evolution can be quite rapid<sup>88</sup>, and the order of serpentine steps can be reversed: a change in a conserved trait facilitates a subsequent adaptive step<sup>59,89</sup>. Fitness trade-offs between conserved and adaptive traits are a common feature of the systems discussed here, and in some cases, the resulting serpentine adaptive dynamics have been observed. For example, resistance or antigenic mutations often have deleterious effects on the folding stability or other conserved functions of a protein<sup>54,57–61,71</sup> and the primary T cell response of HIV virus confers a fitness cost in its Gag p17 protein<sup>16</sup>.

A crucial and largely unexplored determinant of predictability is the variation in initial conditions and environmental factors across





**Figure 3 | From fitness models to evolutionary predictions.** **a**, Minimal fitness seascape of an evolving pathogen: fitness depends on an adaptive trait (resistance or immunity against host defence) and on a conserved trait (protein stability, antigenicity). For concreteness, we assume the seascape has a single moving peak (blue dashed line), which is shown in four consecutive snapshots. Evolution in such seascapes takes serpentine paths: adaptive mutations advance resistance or immunity but have deleterious effects on the conserved trait; compensatory mutations repair these effects. These paths are shown for two lineages in competing clades (solid arrows, past; dashed arrows, future). The time-dependent fitness of each lineage depends on its distance from the peak (green and red dashed lines show trait distance components at a given time  $t$ ). **b, c**, Key steps of predictive analysis. **b**, Fitness models provide estimates of time-dependent phenotypic and fitness differences between competing clades (green, fitness component of the adaptive trait; red, fitness component of the conserved trait; black, total average fitness of a clade at time  $t$ ; bars show fitness cost compared to the peak value marked by a dashed blue line). **c**, Predictions of clade frequency paths into the future of the population, starting at time  $t$  (dashed lines, the blue line shows the combined frequency of all clades other than 1 and 2). In the example, clade 1 increases initially because it has the highest average fitness at time  $t$ ; clade 2 takes over at later times because it harbours a high-fitness subclade.

populations. Many lab evolution experiments are designed to limit this variation: populations start from a well-defined initial state (often a single clone), and the experiments are conducted under carefully calibrated conditions<sup>37,49</sup>. In contrast, the evolution of populations in the wild can have different—and often unknown—initial states, and it takes place under variable ecological conditions. These factors can clearly hamper predictability. However, some recent results indicate that more complex evolutionary processes retain repeatable characteristics. First, standing variation maintains or even enhances short-term repeatability, because adaptive mutations may already be present in the initial population state<sup>28,90</sup>. If selection is sufficiently strong, even de-novo mutations generated from a complex initial state have repeatable features<sup>91</sup>. Second, heterogeneity across parallel-evolving populations can become subdominant if strong adaptive pressure generates convergent evolution. For example, an adaptation experiment of bacteria in the ecosystem of the mouse gut shows similar early-stage phenotypic changes across different hosts<sup>92</sup>. Another case in point is the adaptive immune response of humans to an influenza infection or vaccination. Although individuals have different immune repertoire-wide responses<sup>93</sup>, some antigenic characteristics of their response to related viruses are similar<sup>6</sup>. Despite these convergent aspects, populations are often shaped by differential response to environmental variation. In many cases, this requires modelling evolution under time-dependent selection, in so-called fitness seascapes<sup>94</sup>. The resulting challenges for predictive analysis will be discussed below.

### Predictive data and models

Recent work has underscored the importance of comprehensive data and quantitative modelling for predictions. With modern sequencing, evolutionary models can be based on copious sequence information. We can track the genetic history of entire populations (Fig. 1), detect low-frequency variants, and resolve the spatio-temporal evolutionary dynamics in extended populations<sup>95–97</sup>. An increasingly important direction is to combine sequencing with high-throughput phenotypic and fitness assays<sup>98</sup>. For example, specific interactions between antigens and immune receptors can be inferred by deep mutational scans<sup>99,100</sup>, similar scans map regulatory DNA–protein interactions<sup>98,101–103</sup>.

To build a predictive analysis from these data, we need to relate genetic or phenotypic data to fitness differences in a population. Given the complexity of generic fitness landscapes, this seems a daunting task<sup>77</sup>. Densely sampled sequence data, however, contain copious information on selective effects that can be assembled to infer fitness land- and seascapes. Site-specific amino acid preferences can be inferred from deep sequencing data<sup>104</sup> using equilibrium models of molecular evolution<sup>105,106</sup>; related methods map epistatic interactions between these sites<sup>16,107</sup>. Alternatively, we can infer selection on genetic clades and build predictive models from the local shape of sequence-based coalescent trees<sup>15</sup>.

At the level of quantitative traits, biophysical principles provide powerful guidance for building empirical fitness models<sup>54,58,108–113</sup>. A ubiquitous biophysical trait is the free energy difference  $\Delta G$  between the folded and the unfolded state of a protein, which determines the fraction of successful folds in thermodynamic equilibrium under given physiological conditions. Many other traits depend on the binding of proteins to a molecular target, which is governed by the free energy difference  $\Delta G$  between the bound and the unbound state. Important examples are host–pathogen interactions via binding of immune antibodies to antigenic epitopes; similarly, cancer cells develop peptides presented on the cell surface (called neoantigens) that can be bound by immune T cells<sup>114</sup>. In all such cases, the fitness landscape depending on  $\Delta G$  is strongly constrained by the thermodynamics of the system. These landscapes often take a characteristic ‘mesa’ form<sup>58,62,109</sup> with a plateau at high folding or binding probability, a rapid change of fitness around a characteristic  $\Delta G$  value, and a second plateau at low folding or binding probability. For regulatory or metabolic interactions, additional stabilizing selection against strong binding can modify the landscape to a single fitness peak at intermediate values of  $\Delta G$ . Importantly, such landscapes have few fit parameters that can be learned from training data<sup>62,80</sup>.

A minimal fitness model for pathogen evolution can serve to illustrate key concepts of predictive analysis. The model describes the coupled evolution of an adaptive trait (such as antigenicity or resistivity) and a conserved trait (for example, fold stability), which are encoded in a single protein. The minimal fitness seascape, which contains stabilizing selection on the conserved trait and adaptive pressure on the adaptive trait<sup>94</sup>, is an explicitly time-dependent

version of Fisher's geometrical model<sup>115,116</sup> (Fig. 3a) or a similar model with components of mesa form (as described above). This type of model has been applied to the evolution of human influenza<sup>13</sup>; its fitness trade-off between traits also captures aspects of HIV evolution under host immune pressure<sup>17</sup>, of drug resistance evolution<sup>61</sup>, and of cancer evolution<sup>63</sup>. The time-dependence of selection on the adaptive trait describes variable environments and is a key feature of the model. For example, the cross-immunity interactions affecting a pathogen<sup>13</sup> depend on the infection history of its hosts; similarly, movement along a spatial gradient of drug concentration generates time-dependent adaptive pressure<sup>117,118</sup>.

How can we gather data to inform such fitness models? We often have at least partial information on the genetic changes underlying the evolution of the relevant phenotypes. In viral pathogens, for example, the antigenic evolution predominantly occurs in specific epitope sites, whereas fold stability has a broader mutational target of amino acid changes throughout the protein. In some cases, the genetic information includes epistatic interactions between specific sequence sites<sup>16,107</sup>. Predictive analysis can then exploit an approximate genotype–phenotype map of the adaptive process. Alternatively, we can record phenotypic data by experiment. For example, antigenic assays<sup>119–121</sup> or time-resolved deep sequencing of immune repertoires can track the selection pressures that drive the adaptive evolution of a pathogen population. We can then use a phenotypic fitness model, such as the minimal seascape, to compute fitness differences between strains in a population (Fig. 3b). In this way, we can predict frequency trajectories of competing clades<sup>13</sup> (Fig. 3c) and estimate the probability of future escape mutations that evade host immunity<sup>16</sup>.

From these examples, we can distil a few general lessons for predictive modelling in evolution. A 'mechanistic' fitness model of the evolutionary dynamics, similar to our minimal model, is feasible if the population harbours substantial variation in fitness that can be explained by few key phenotypes. Such models generically contain positive and negative fitness components, which jointly constrain the evolutionary complexity of the system and generate predictability. In adaptive processes, modelling starts with the key adaptive traits of the system, such as antibiotic resistance or immunity against an antigen. Importantly, however, an adaptive trait alone is often an insufficient basis for predictions, because its evolution is generically coupled to other traits. As discussed above, such correlations arise from epistasis or pleiotropy and generate a serpentine pattern of adaptive paths (Fig. 3a). They can reduce the independent components of fitness variation and, thus, reduce the necessary complexity of fitness models. In complex organisms, we need to map the most informative phenotypes and their correlations to determine the normal modes of predictive analysis. This will eventually require a systems-biology approach to evolutionary predictions<sup>122</sup>.

The above examples also show that understanding the ecology of fast-evolving populations, which includes exposure to drugs and host–pathogen interactions, is often the salient point of predictive analysis. Co-evolutionary fitness models, which have recently been developed for pathogen–immune systems<sup>123,124</sup>, are a promising step towards predictions in realistic ecological settings. The success of these models will depend on sufficiently dense time-resolved data of the evolving population and its variable environment. Predicting evolution in an ecological context also generates new questions. For example, we often want to predict not only frequencies but absolute population numbers, such as the viral load of an infection, the size of a cancer cell population, or the size of an epidemic<sup>125</sup>. These numbers depend on absolute fitness values, which in turn respond strongly to ecological determinants of reproduction. Moreover, in heterogeneous populations of fast-evolving systems, fitness differences within a population can be of the same order of magnitude as absolute growth rates, so population size dynamics must be modelled together with the evolution of clade frequencies. This problem

is difficult in general, but at least the response of pathogen population size to immune or vaccination pressure can be computed using fitness models of immune interactions<sup>13,120,121</sup>. Maximizing this response has been exploited as a criterion for influenza vaccine strain selection<sup>13</sup>.

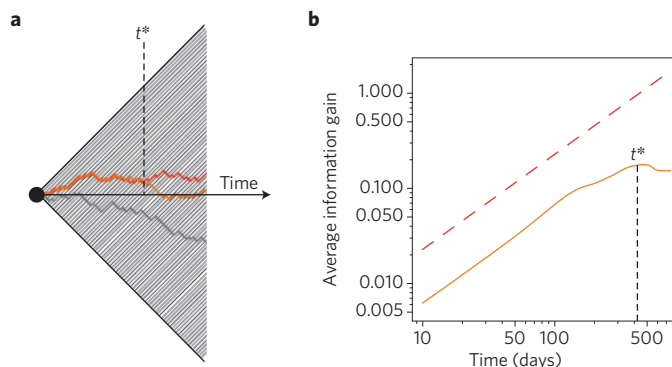
An exciting complement to *in silico* modelling is to use laboratory evolution for predicting an evolutionary process in the wild. This makes sense if we can find a laboratory model that evolves faster than the primary system or can be run in multiple replicates<sup>126</sup>. For example, massively parallel tumour cell cultures can reveal likely future resistance mutations<sup>127</sup>. Once these methods can be applied efficiently to individual tumours, they may circumvent the problem of genetic uniqueness and provide patient-specific predictions of tumour response to therapy. Clearly, the increasing knowledge on parallelization and replicability of laboratory evolution will prove very useful for the design of such assays.

### Measuring prediction quality

Predictive analysis will be applied to a broadening range of systems, and it will be built on increasingly diverse data and methods. To keep a critical eye on quality, we need an unbiased way to gauge predictive success. Intuitively, we have an idea of what makes a good prediction: it has an element of surprise and an element of truth. To illustrate these criteria, suppose we conduct an evolution experiment with mice and assert the outcome of this experiment will be mice with four legs. This statement is likely to be true but unsurprising; most would rate it as obvious in the first place. On the other hand, we may predict the experiment to produce five-legged mice. That statement is quite surprising but, as performing the actual experiment would show, is unlikely to hold up to testing. The example demonstrates that any prediction is a probabilistic statement. Specifically, it is a bet about the future that should strike a balance between surprise and truth.

We can use information theory to quantify these criteria: good predictions combine low probability in our prior expectation (that is, they are surprising) and high predicted probability (that is, they come close to the truth) of the actual process as observed later. The 'information gain', defined as the log ratio of predicted and prior probability, measures how much the prediction reduces our uncertainty about the future process<sup>13,128</sup>. For an adaptive process, the information gain is closely related to the amount of adaptation (that is, the cumulative fitness flux<sup>129</sup>) explained by the prediction model. Figure 4a illustrates how prior and posterior probability of a prediction depend on the evolutionary paths of the system and on the time interval of predictions. The prior probability of any future evolutionary path rapidly decreases with time, because longer processes have a much higher number of *a priori* plausible paths than shorter ones. Specifying the prior probability requires a statistical null model (for example, a neutral model assigns equal probability to all paths of the same length). For a good prediction, the actual path remains likely for some time, but its probability must eventually decay because of noise in the data and imperfections of the model. Therefore, the information gain shows an initial increase and saturates at a characteristic time. This sets the 'time horizon' of the prediction method, beyond which the results cannot be trusted.

As an example, Fig. 4b shows the information gain of evolutionary predictions for the human influenza virus A/H3N2. We predict the evolutionary path of clade frequencies by an antigenicity–stability fitness model as described above and evaluate the information gain of these predictions compared to a null model of neutral evolution<sup>13</sup> (M. Łuksza and M. Lässig, manuscript in preparation). As shown by the time-dependence of the information gain, the model predictions capture the actual evolutionary process with a time horizon in the order of one year. How much this horizon can be extended by improved modelling remains an open question.



**Figure 4 | Information gain and time horizon of predictions.** **a**, Prior and posterior probabilities of evolutionary paths (schematic). The path of the actual process (red) is shown together with the most likely predicted path (orange) and a random alternative path of equal prior probability (grey). The prior probability of any path rapidly decreases with time (in a neutral null model,  $p(t) = 2^{-t}$ , because this example has two possible directions at each grid point). The prediction initially keeps up with the actual process—that is, the predicted probability  $q(t)$  of the actual path remains high—but becomes random beyond the time horizon  $t^*$ . **b**, Information gain of evolutionary predictions for human influenza A/H3N2. For the observed evolutionary path of clade frequencies (Fig. 1b), we evaluate the prior probability under a neutral null model,  $p(t)$ , and the predicted probability under the antigenicity–stability fitness model<sup>13</sup> (M. Łuksza and M. Lässig, manuscript in preparation),  $q(t)$  (see section ‘Predictive data and models’). The average information gain  $I(t) = \log[q(t)/p(t)]$  of the model prediction (orange line) is positive, which shows that the fitness model is a better predictor than the null model. We also plot the corresponding information gain of posterior tracking, which uses fitness values inferred from frequency changes in the actual process (red dashed line). Averages are over yearly predictions in the period 2004–16; the prediction starts at the end of February in each year. The prediction scheme is seen to capture the actual process up to a time horizon  $t^*$  of order one year. Figure created by M. Łuksza.

The link of evolutionary predictions to information theory underscores an important general point: the predictability of an evolutionary process is not a yes-or-no issue, but is itself a quantitative trait. We can probe this trait by the information gain of actual predictions, which can be evaluated by comparison with posterior data. In this way, we can compare the predictability of different evolutionary processes by a given method, as well as the prediction quality of different methods for a given process.

### From prediction to control

Any therapy or intervention against a fast-evolving pathogen is an attempt to control its future population. Such interventions have different goals and strategies, which range from controlling an infection or cancer within an individual patient to reducing the global spread of pathogen resistance<sup>130,131</sup>. Similarly, the adaptive immune system can be seen as a host’s intrinsic strategy to control pathogens<sup>133</sup>. There is a fundamental link between predicting an evolutionary process and controlling its future outcome. This is because a predictive computational or experimental model does not just reproduce the actual process; it generates an entire probability distribution of possible outcomes. Controlling the process amounts to changing that distribution by means of an external evolutionary pressure. The intended change is often drastic: an *a priori* likely outcome, such as the occurrence of resistance mutations or the increase of pathogen load, is to become unlikely. If our intervention or therapy can produce the required evolutionary pressure, predictive models can be leveraged to nudge the process towards the intended outcome. Specifically, we can include the control as an

additional component into a fitness model and evaluate the evolutionary response of the population to a given control protocol. For example, a xenograft mouse model of melanoma shows increased survival when the drug is withdrawn at predefined time points<sup>132</sup>, and fitness modelling of these dynamics predicts how the drug protocol can be optimized based on real-time measurements to further increase survival<sup>134</sup>. HIV combination therapy, a protocol of multiple suppressive drugs, is a classic case of evolutionary control aimed minimizing the rate of viral escape mutations<sup>135</sup>. Similarly, a successful vaccine against HIV needs to trigger an immune response that co-evolves with the virus<sup>123</sup>.

Evolutionary control can reinforce itself if the external adaptive pressure enhances predictability by constraining evolutionary paths (Fig. 2b). For instance, melanoma cells carrying a given mutation in the *BRAF* oncogene show strong initial response to the drug vemurafenib<sup>136</sup>, but most cancers of this kind will eventually relapse. The escape to drug resistance appears to be via few mutational pathways, which can be used for predictive analysis of second-line therapy choices. In the coming years, we will have increasingly detailed and time-resolved data of evolutionary pressure and response, for example on immune response to infections by antigen-specific and broadly neutralizing antibodies<sup>137</sup>. Combined with co-evolutionary fitness models<sup>120,121</sup> and fitness models of metabolic pathways under stress<sup>138</sup>, such data will open new avenues of designing and optimizing evolutionary control.

Prediction and control based on mechanistic evolutionary models are always imperfect, because our knowledge of population data and dynamical parameters remains incomplete for even the simplest biological systems. It is useful to compare mechanistic models with model-free methods, such as deep reinforcement learning. Recent studies have presented remarkable model-free solutions of complex problems; for example, computers can learn to play video games without *a priori* knowledge of the game<sup>139</sup>. Can we control an evolving population in a similar way, without prior knowledge of the evolutionary rules? This is far from obvious, given substantial differences in data structure and learning dynamics. Computer game records are comprehensive, free, and fast to acquire; in contrast, evolutionary data are always incomplete, comparatively costly, and ‘computing’ by evolutionary processes is slow. These differences may favour simplified mechanistic models as an avenue to successful prediction and control of evolution.

The link between prediction and control is crucial for ethically responsible decision-making. For example, judging genome editing manipulations must include the question of how predictable their outcome is. Our discussion of evolutionary correlations between phenotypes shows how complex this task is: we have to assess the primary effect of the manipulation, but also secondary changes in other traits that are generated by pleiotropy and epistasis (Fig. 3a). We also need to gauge the effects and persistence of genetic changes under changing environmental and co-evolutionary conditions. For example, drug-resistance mutations can sometimes remain fixed in a population through subsequent epistatic mutations, even when the drug is no longer present and the resistance mechanism bears a fitness cost<sup>140</sup>. In all of these systems, predictive modelling will take an important role in designing responsible control strategies.

### Conclusion

For a growing number of systems, we are witnessing the transition to a new kind of predictive evolutionary biology. In this Perspective, we focused on a specific mode of evolution: fast, predominantly asexual processes driven by a large supply of mutations and strong selection. That is a promising starting point for predictive analysis, but the spectrum of modes and time scales in evolution is clearly much broader. Work in the years ahead will show how predictability plays out in more complex systems, including populations with various rates of recombination. Some of these concepts may also be



extendable to repeated patterns in the macro-evolution of multicellular organisms. We expect that the endeavour of predictive analysis will affect our overall view of the life sciences. It will provide a rational basis for decision-making in a number of areas of medicine and public health. At a more fundamental level, it will promote a unifying view on different organisms based on common dynamical principles. Optimizing predictions is a way to learn what the evolutionarily relevant functions of the system are: biology informs predictions and predictions inform biology.

Received 3 May 2016; accepted 10 January 2017;  
published 21 February 2017

## References

- Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (1989).
- Orgogozo, V. Replaying the tape of life in the twenty-first century. *Interface Focus* **5**, 20150057 (2015).
- Lobkovsky, A. E. & Koonin, E. V. Replaying the tape of life: quantification of the predictability of evolution. *Front. Genet.* **3**, 246 (2012).
- Achaz, G., Rodríguez-Verdugo, A., Gaut, B. S. & Tenaillon, O. The reproducibility of adaptation in the light of experimental evolution with whole genome sequencing. *Adv. Exp. Med. Biol.* **781**, 211–231 (2014).
- Salverda, M. L. M. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7**, e1001321 (2011).
- Fonville, J. M. *et al.* Antibody landscapes after influenza virus infection or vaccination. *Science* **346**, 996–1000 (2014).
- Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
- Lipinski, K. A. *et al.* Cancer evolution and the limits of predictability in precision cancer medicine. *Trends Cancer* **2**, 49–63 (2016).
- Greaves, M. Evolutionary determinants of cancer. *Cancer Discov.* **5**, 806–820 (2015).
- Lieberman, T. & Michel, J. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275–1280 (2011).
- Miles, J. J., Douek, D. C. & Price, D. A. Bias in the  $\alpha\beta$  T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375–387 (2011).
- Bull, J. J. & Molineux, I. J. Predicting evolution from genomics: experimental evolution of bacteriophage T7. *Heredity (Edinb.)* **100**, 453–463 (2008).
- Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
- Munck, C., Gumpert, H. K., Wallin, A. I. N., Wang, H. H. & Sommer, M. O. A. Prediction of resistance development against drug combinations by collateral responses to component drugs. *Sci. Transl. Med.* **6**, 262ra156 (2014).
- Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *eLife* **3**, e03568 (2014).
- Barton, J. P. *et al.* Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat. Commun.* **7**, 11660 (2016).
- Neher, R. A. Genetic drift, selective interference, and population genetics of rapid adaptation. *Annu. Rev. Ecol. Syst.* **44**, 195–215 (2013).
- Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).
- Miralles, R., Gerrish, P., Moya, A. & Elena, S. Clonal interference and the evolution of RNA viruses. *Science* **285**, 1745–1747 (1999).
- Lang, G. I., Botstein, D. & Desai, M. M. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* **188**, 647–661 (2011).
- Strelkova, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192**, 671–682 (2012).
- Miller, J. D. *et al.* Human effector and memory CD8<sup>+</sup> T cell responses to smallpox and yellow fever vaccines. *Immunity* **28**, 710–722 (2008).
- Schuh, A., Becq, J., Humphray, S. & Alexa, A. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191–4197 (2012).
- Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
- Hoehn, K. B. *et al.* Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Philos. Trans. R. Soc. B* **370**, 20140241 (2015).
- Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
- Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
- Long, A., Liti, G., Luptak, A. & Tenaillon, O. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat. Rev. Genet.* **16**, 567–582 (2015).
- Lang, G. I. & Desai, M. M. The spectrum of adaptive mutations in experimental evolution. *Genomics* **104**, 412–416 (2014).
- Stratton, M., Campbell, P. & Futreal, P. The cancer genome. *Nature* **458**, 719–724 (2009).
- Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
- Venturi, V., Price, D. A., Douek, D. C. & Davenport, M. P. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
- Mora, T. & Walczak, A. M. Quantifying lymphocyte receptor diversity. Preprint at *bioRxiv* <https://doi.org/10.1101/046870> (2016).
- Elhanati, Y. *et al.* Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. B* **370**, 20140243 (2015).
- Thomas, N. *et al.* Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30**, 3181–3188 (2014).
- Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9**, e1003972 (2013).
- Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839 (2013).
- Kryazhimskiy, S., Rice, D. P., Jerison, E. R. & Desai, M. M. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522 (2014).
- Toprak, E. *et al.* Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.* **44**, 101–105 (2012).
- Barroso-Batista, J., Demengeot, J. & Gordo, I. Adaptive immunity increases the pace and predictability of evolutionary change in commensal gut bacteria. *Nat. Commun.* **6**, 8945 (2015).
- Christin, P.-A., Weinreich, D. M. & Besnard, G. Causes and evolutionary significance of genetic convergence. *Trends Genet.* **26**, 400–405 (2010).
- Ramiro, R. S., Costa, H. & Gordo, I. Macrophage adaptation leads to parallel evolution of genetically diverse *Escherichia coli* small-colony variants with increased fitness *in vivo* and antibiotic collateral sensitivity. *Evol. Appl.* **9**, 994–1004 (2016).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Gerstung, M. *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* **6**, 5901 (2014).
- Murali-Krishna, K. *et al.* Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. *Immunity* **8**, 177–187 (1998).
- Butler, T. C., Kardar, M. & Chakraborty, A. K. Quorum sensing allows T cells to discriminate between self and nonself. *Proc. Natl Acad. Sci. USA* **110**, 11833–11838 (2013).
- Schiffels, S., Szölösi, G. J., Mustonen, V. & Lässig, M. Emergent neutrality in adaptive asexual evolution. *Genetics* **189**, 1361–75 (2011).
- Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109**, 4950–4955 (2012).
- Elena, S. F. & Lenski, R. E. Microbial genetics: evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
- Betancourt, A. J. Genomewide patterns of substitution in adaptively evolving populations of the RNA bacteriophage MS2. *Genetics* **181**, 1535–1544 (2009).
- Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F. & Godelle, B. Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics* **152**, 485–493 (1999).
- Illingworth, C. J. R. & Mustonen, V. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLoS Pathog.* **8**, e1003091 (2012).
- Nourmohammad, A., Held, T. & Lässig, M. Universality and predictability in molecular quantitative genetics. *Curr. Opin. Genet. Dev.* **23**, 684–693 (2013).
- Zeldovich, K. B., Chen, P. & Shakhnovich, E. I. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl Acad. Sci. USA* **104**, 16152–16157 (2007).
- Marcusson, L. L., Frimodt-Møller, N. & Hughes, D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog.* **5**, e1000541 (2009).
- MacLean, R. C., Hall, A. R., Perron, G. G. & Buckling, A. The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. *Nat. Rev. Genet.* **11**, 405–414 (2010).
- Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–71 (2010).
- Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl Acad. Sci. USA* **108**, 9916–9921 (2011).

59. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).
60. Rodrigues, J. V. *et al.* Biophysical principles predict fitness landscapes of drug resistance. *Proc. Natl Acad. Sci. USA* **113**, E1470–E1478 (2016).
61. Hughes, D. & Andersson, D. I. Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms. *Nat. Rev. Genet.* **16**, 459–471 (2015).
62. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl Acad. Sci. USA* **110**, 13067–13072 (2013).
63. McFarland, C. D. *et al.* Passenger DNA alterations reduce cancer fitness in cell culture and mouse models. Preprint at *bioRxiv* <https://dx.doi.org/10.1101/026302> (2015).
64. McFarland, C. D., Mirny, L. A. & Korolev, K. S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc. Natl Acad. Sci. USA* **111**, 15138–15143 (2014).
65. Koelle, K. & Rasmussen, D. A. The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *eLife* **4**, e07361 (2015).
66. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
67. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106–110 (2011).
68. Palmer, A. C. & Kishony, R. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nat. Rev. Genet.* **14**, 243–248 (2013).
69. Rodríguez-Verdugo, A., Gaut, B. S. & Tenaillon, O. Evolution of *Escherichia coli* rifampicin resistance in an antibiotic-free environment during thermal stress. *BMC Evol. Biol.* **13**, 50 (2013).
70. Weinreich, D., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
71. Schenk, M. F. & de Visser, J. A. G. M. Predicting the evolution of antibiotic resistance. *BMC Biol.* **11**, 14 (2013).
72. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
73. McKeown, A. N. *et al.* Evolution of DNA Specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58–68 (2014).
74. Couce, A. & Tenaillon, O. A. The rule of declining adaptability in microbial evolution experiments. *Front. Genet.* **6**, 99 (2015).
75. Press, M. O., Queitsch, C. & Borenstein, E. Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.* **26**, 826–833 (2016).
76. Heckmann, D. *et al.* Predicting  $C_4$  photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell* **153**, 1579–1588 (2013).
77. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
78. Bank, C., Matuszewski, S., Hietpas, R. T. & Jensen, J. D. On the (un) predictability of a large intragenic fitness landscape. *Proc. Natl Acad. Sci. USA* **113**, 14085–14090 (2016).
79. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
80. du Plessis, L., Leventhal, G. & Bonhoeffer, S. How good are statistical models at approximating complex fitness landscapes. *Mol. Biol. Evol.* **33**, 2454–2468 (2016).
81. Levin, B. R., Perrot, V. & Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154**, 985–997 (2000).
82. Luciani, F., Sisson, S. A., Jiang, H., Francis, A. R. & Tanaka, M. M. The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **106**, 14711–14715 (2009).
83. Vogwill, T. & MacLean, R. C. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evol. Appl.* **8**, 284–295 (2015).
84. Reynolds, M. G. Compensatory evolution in rifampin-resistant *Escherichia coli*. *Genetics* **156**, 1471–1481 (2000).
85. Brown, K. M., Depristo, M. A., Weinreich, D. M. & Hartl, D. L. Temporal constraints on the incorporation of regulatory mutants in evolutionary pathways. *Mol. Biol. Evol.* **26**, 2455–2462 (2009).
86. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage X174. *Genetics* **170**, 989–999 (2005).
87. Maisnier-Patin, S., Berg, O. G., Liljas, L. & Andersson, D. I. Compensatory adaptation to the deleterious effect of antibiotic resistance in *Salmonella typhimurium*. *Mol. Microbiol.* **46**, 355–366 (2002).
88. Sousa, A., Magalhães, S. & Gordo, I. Cost of antibiotic resistance and the geometry of adaptation. *Mol. Biol. Evol.* **29**, 1417–1428 (2012).
89. Aakre, C. D. *et al.* Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
90. Burke, M. K., Liti, G. & Long, A. D. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. *Mol. Biol. Evol.* **31**, 3228–3239 (2014).
91. Vázquez-García, I. *et al.* Background-dependent effects of selection on subclonal heterogeneity. Preprint at *bioRxiv* <https://dx.doi.org/10.1101/039859> (2016).
92. Barroso-Batista, J. *et al.* The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet.* **10**, e1004182 (2014).
93. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra19 (2013).
94. Mustonen, V. & Lässig, M. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.* **25**, 111–119 (2009).
95. Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl Acad. Sci. USA* **104**, 19926–19930 (2007).
96. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170 (2016).
97. Levy, S. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
98. Venkataram, S. *et al.* Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. *Cell* **166**, 1585–1589.e22 (2016).
99. Adams, R., Kinney, J. B., Mora, T. & Walczak, A. M. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* <http://dx.doi.org/10.7554/eLife.23156> (2016).
100. Birnbaum, M. E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).
101. Maerkl, S. & Quake, S. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
102. Bulyk, M. DNA microarray technologies for measuring protein–DNA interactions. *Curr. Opin. Biotechnol.* **17**, 422–430 (2006).
103. Kinney, J., Murugan, A., Callan Curtis, G. J. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA* **107**, 9158–9163 (2010).
104. Thyagarajan, B. & Bloom, J. D. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* **3**, e03300 (2014).
105. Halpern, A. L. & Bruno, W. J. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15**, 910–917 (1992).
106. Berg, J., Willmann, S. & Lässig, M. Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.* **4**, 42 (2004).
107. Kouyos, R. D. *et al.* Exploring the complexity of the HIV-1 fitness landscape. *PLoS Genet.* **8**, e1002551 (2012).
108. Lässig, M. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinform.* **8**(suppl. 6), S7 (2007).
109. Mustonen, V., Kinney, J. B., Callan Curtis, G. J. & Lässig, M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl Acad. Sci. USA* **105**, 12376–12381 (2008).
110. Serohijos, A. W. R. & Shakhnovich, E. I. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* **26**, 84–91 (2014).
111. Moses, A., Chiang, D., Pollard, D., Iyer, V. & Eisen, M. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome* **5**, R98 (2004).
112. Liberles, D. A. *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **21**, 769–785 (2012).
113. Gerland, U. & Hwa, T. On the selection and evolution of regulatory DNA motifs. *J. Mol. Evol.* **55**, 386–400 (2002).
114. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
115. Fisher, R. A. *The Genetical Theory of Natural Selection*. (Clarendon, 1930).
116. Trindade, S., Sousa, A. & Gordo, I. Antibiotic resistance and stress in the light of Fisher's model. *Evolution* **66**, 3815–3824 (2012).
117. Hermesen, R., Deris, J. B. & Hwa, T. On the rapidity of antibiotic resistance evolution facilitated by a concentration gradient. *Proc. Natl Acad. Sci. USA* **109**, 10775–10780 (2012).
118. Greulich, P., Waclaw, B. & Allen, R. J. Mutational pathway determines whether drug gradients accelerate evolution of drug-resistant cells. *Phys. Rev. Lett.* **109**, 88101 (2012).
119. Smith, D. J., Lapedes, A. S. & De Jong, J. C. Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–377 (2004).
120. Bedford, T. *et al.* Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3**, e01914 (2014).
121. Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A. & Shraiman, B. I. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl Acad. Sci. USA* **113**, E1701–E1709 (2016).



122. Papp, B., Notebaart, R. A. & Pál, C. Systems-biology approaches for predicting genomic evolution. *Nat. Rev. Genet.* **12**, 591–602 (2011).
123. Wang, S. *et al.* Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* **160**, 785–797 (2015).
124. Nourmohammad, A., Otwinowski, J. & Plotkin, J. B. Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genet.* **12**, e1006171 (2016).
125. Gandon, S., Day, T., Metcalf, C. J. E. & Grenfell, B. T. Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends Ecol. Evol.* **31**, 1–13 (2016).
126. Torres-Barceló, C. & Hochberg, M. E. Evolutionary rationale for phages as complements of antibiotics. *Trends Microbiol.* **24**, 249–256 (2016).
127. Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
128. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
129. Mustonen, V. & Lässig, M. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl Acad. Sci. USA* **107**, 4248–4253 (2010).
130. Read, A. F., Day, T. & Huijben, S. The evolution of drug resistance and the curious orthodoxy of aggressive chemotherapy. *Proc. Natl Acad. Sci. USA* **108**, 10871–10877 (2011).
131. Ankomah, P. & Levin, B. R. Exploring the collaboration between antibiotics and the immune response in the treatment of acute, self-limiting infections. *Proc. Natl Acad. Sci. USA* **111**, 8331–8338 (2014).
132. Das Thakur, M. *et al.* Modelling vemurafenib resistance in melanoma reveals a strategy to forestall drug resistance. *Nature* **494**, 251–255 (2013).
133. Mayer, A., Mora, T., Rivoire, O. & Walczak, A. M. Diversity of immune strategies explained by adaptation to pathogen statistics. *Proc. Natl Acad. Sci. USA* **113**, 8630–8635 (2016).
134. Fischer, A., Vázquez-García, I. & Mustonen, V. The value of monitoring to control evolving populations. *Proc. Natl Acad. Sci. USA* **112**, 1007–1012 (2015).
135. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586 (1996).
136. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
137. Caskey, M. *et al.* Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
138. Perfeito, L., Ghozzi, S., Berg, J., Schnetz, K. & Lässig, M. Nonlinear fitness landscape of a molecular pathway. *PLoS Genet.* **7**, e1002160 (2011).
139. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
140. Abel Zur Wiesch, P., Kouyos, R., Engelstädter, J., Regoes, R. R. & Bonhoeffer, S. Population biological principles of drug-resistance evolution in infectious diseases. *Lancet Infect. Dis.* **11**, 236–247 (2011).
141. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).
142. Liao, H.-X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).

## Acknowledgements

We thank M. Desai, I. Gordo, M. Łuksza, T. Mora and A. Nourmohammad for comments on the manuscript. M. Desai, M. Łuksza and A. Nourmohammad also provided important input to illustrations. This work has been partially supported by Deutsche Forschungs-gemeinschaft grant SFB 680 (M.L.), Wellcome Trust grant 098051 (V.M.), and European Research Council ERCStG 306312 (A.M.W.).

## Author contributions

All authors developed concepts and wrote the paper. Authors are listed in alphabetical order.

## Additional information

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

Correspondence should be addressed to M.L., V.M. and A.M.W.

**How to cite this article:** Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nat. Ecol. Evol.* **1**, 0077 (2017).

## Competing interests

The authors declare no competing financial interests.