PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Rajon E, Plotkin JB. 2013 The evolution of genetic architectures underlying quantitative traits. Proc R Soc B 280: 20131552.
http://dx.doi.org/10.1098/rspb.2013.1552

Received: 14 June 2013
Accepted: 1 August 2013

Subject Areas:
evolution, genetics, theoretical biology

Keywords:
genetic architecture, quantitative trait loci, empirical quantitative trait loci, copy number, polygenic

Author for correspondence:
Etienne Rajon
e-mail: rajon@sas.upenn.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2013.1552 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

The evolution of genetic architectures underlying quantitative traits

Etienne Rajon and Joshua B. Plotkin

Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

In the classic view introduced by R. A. Fisher, a quantitative trait is encoded by many loci with small, additive effects. Recent advances in quantitative trait loci mapping have begun to elucidate the genetic architectures underlying vast numbers of phenotypes across diverse taxa, producing observations that sometimes contrast with Fisher's blueprint. Despite these considerable empirical efforts to map the genetic determinants of traits, it remains poorly understood how the genetic architecture of a trait should evolve, or how it depends on the selection pressures on the trait. Here, we develop a simple, population-genetic model for the evolution of genetic architectures. Our model predicts that traits under moderate selection should be encoded by many loci with highly variable effects, whereas traits under either weak or strong selection should be encoded by relatively few loci. We compare these theoretical predictions with qualitative trends in the genetics of human traits, and with systematic data on the genetics of gene expression levels in yeast. Our analysis provides an evolutionary explanation for broad empirical patterns in the genetic basis for traits, and it introduces a single framework that unifies the diversity of observed genetic architectures, ranging from Mendelian to Fisherian.

1. Introduction

A quantitative trait is encoded by a set of genetic loci whose alleles contribute directly the trait value, interact epistatically to modulate each other's contributions, and possibly contribute to other traits. The resulting genetic architecture of a trait [1] influences its variational properties [2–5], and therefore affects a population's capacity to adapt to new environmental conditions [1,6,7]. Over longer time scales, genetic architectures of traits have important consequences for the evolution of recombination [8] and of sex [9], and even reproductive isolation and speciation [10].

Although scientists have studied the genetic basis for phenotypic variation for more than a century, recent technologies, as well as the promise of agricultural and medical applications, have stimulated tremendous efforts to map quantitative trait loci (QTL) in diverse taxa [11–19]. These studies have revealed many traits that seem to rely on Fisherian architectures, with contributions from many loci [20], whose additive effects are often so small that QTL studies lack power to detect them individually [16,21,22]. Other traits, however, are encoded by a relatively small number of loci—including the large number of human phenotypes with known Mendelian inheritance.

The subtle statistical issues of designing and interpreting QTL studies in order to accurately infer the molecular determinants of a trait are already actively studied [16,21,22]. Nevertheless, distinct from these statistical issues of inferences from empirical data, we lack a theoretical framework for forming *a priori* expectations about the genetic architecture underlying a trait [1,4]. For instance, what types of traits should we expect to be monogenic, and what traits should be highly polygenic? More generally, how does the genetic architecture underlying a trait evolve, and what features of a trait shape the evolution of its architecture? To address these questions, we developed a mathematical model for the evolution of genetic architectures, and we compared its predictions with a large body of empirical data on quantitative traits.

2. Results and discussion

(a) Genetic architectures predicted by a population-genetic model

Our approach to understanding the evolution of genetic architectures combines standard models from quantitative genetics [23] with the Wright–Fisher model from population genetics [24]. In its simplest version, our model considers a continuous trait whose value, x, is influenced by L loci. Each locus i contributes additively an amount α_i, so that the trait value is defined as the mean of the α_i values across contributing loci. This trait definition means that a gene's contribution to a trait is diluted when L is large, which prevents direct selection on gene copy numbers when genes have similar contributions [25,26]. We discuss this definition below, along with alternatives such as the sum. The fitness of an individual with trait value x is assumed to be Gaussian with mean zero and standard deviation σ_f, so that smaller values of σ_f correspond to stronger stabilizing selection on the trait [23]. Individuals in a population of size N replicate according to their relative fitnesses. Upon replication, an offspring may acquire a point mutation that alters the direct effect of one locus, i, perturbing the value of α_i for the offspring by a normal deviate; or the offspring may experience a duplication or a deletion in a contributing locus, which changes the number of loci L that control the trait value in that individual (see Methods). Point mutations, duplications and deletions occur at rates μ, r_{dup}, r_{del}, which have comparable magnitudes in nature (see electronic supplementary material, table S1) [27–30]. We assume that deletions are more frequent than duplications, for two reasons. First, in our model, deletions represent both actual deletions and loss-of-function mutations. Second, even ignoring loss-of-function mutations, the frequency of deletions is known to exceed that of duplications [31]. Finally, an offspring may also increase the number of loci that contribute to its trait value by recruitment; that is, by acquiring a recruitment mutation, with probability $\mu \times r_{rec}$, in some gene that did not previously contribute to the trait value (see Methods).

Over successive generations in our model, the genetic architecture underlying the trait (i.e. how many loci contribute to the trait's value and the extent of their contributions) varies among the individuals in the population, and evolves. The genetic architectures that evolve in our model represent the complete genetic determinants of a trait, which may include—but do not correspond precisely to—the genetic loci that would be detected based on polymorphisms segregating in a sample of individuals in a QTL study. We discuss this important distinction below, when we compare the predictions of our model with empirical QTL data.

We studied the evolution of genetic architectures in sets of 500 replicate populations, simulated by Monte Carlo, with different amounts of selection on the trait. We ran each of these simulations for 50 million generations, in order to model the extensive evolutionary divergence over which genetic architectures are assembled in nature. The form of the genetic architecture that evolves in our model depends critically on the strength of selection on the trait. In particular, we found a striking non-monotonic pattern: the equilibrium number of loci that influence a trait is greatest when the strength of selection on the trait is intermediate (figure 1). The distribution of allelic effects also evolves in our model

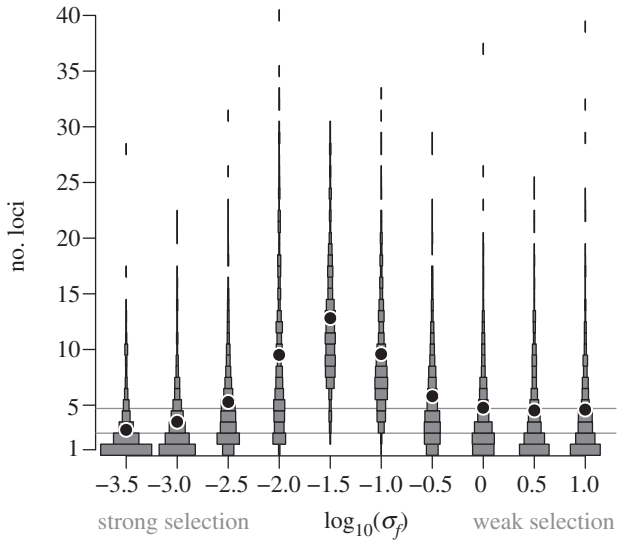

Figure 1. The genetic architecture underlying a trait depends on the strength of selection on the trait, in a population-genetic model. Traits subject to intermediate selection (intermediate values of σ_f) evolve genetic architectures with the greatest number of controlling loci. Dots denote the mean number of loci in the architecture underlying a trait, among 500 replicate Wright–Fisher simulations, for each value of the selection pressure σ_f. The rectangular areas represent the distribution of the number of loci in the architecture. The neutral expectations for the equilibrium number of loci (see Methods) are represented as grey lines, when recruitment events are neutral (top line) or not (bottom line). Parameters are set to their default values (see electronic supplementary material, table S2).

(see electronic supplementary material, figure S1). The variance of allelic effects across loci exhibits a similar pattern as the one observed for the number of contributing loci: it is maximized for traits under intermediate selection. In other words, our model predicts that traits under moderate selection will be encoded by many loci with highly divergent effects, whereas traits under strong or weak selection will be encoded by relatively few loci. Moreover, the distribution of allelic effects determines the typical phenotypic effects of gene deletions and duplications, which are also greatest for traits under intermediate selection (see electronic supplementary material, figure S2).

We also studied how epistatic interactions among loci influence the evolution of genetic architecture. To incorporate the influence of locus j on the contribution of locus i, we introduced epistasis parameters β_{ji}, so that the trait value is now given by

$$x = \frac{1}{L}\sum_{i=1}^{L}\left(\alpha_i \times f_\beta\left(\sum_{j=1}^{L}\beta_{ji}\right)\right), \qquad (2.1)$$

where f_β is a standard sigmoidal filter function (see Methods; see also electronic supplementary material, figure S4*a*) [8]. As with the direct effects of loci, the epistatic effects were allowed to mutate and vary within the population, and evolve. Although significant epistatic interactions emerge in the evolved populations, and modulate the direct contributions of individual alleles (see electronic supplementary material, figure S3*b*,*c*), the presence of epistasis does not strongly affect the average number of loci that control a trait (see electronic supplementary material, figures S3*a* and S4). Epistasis is not

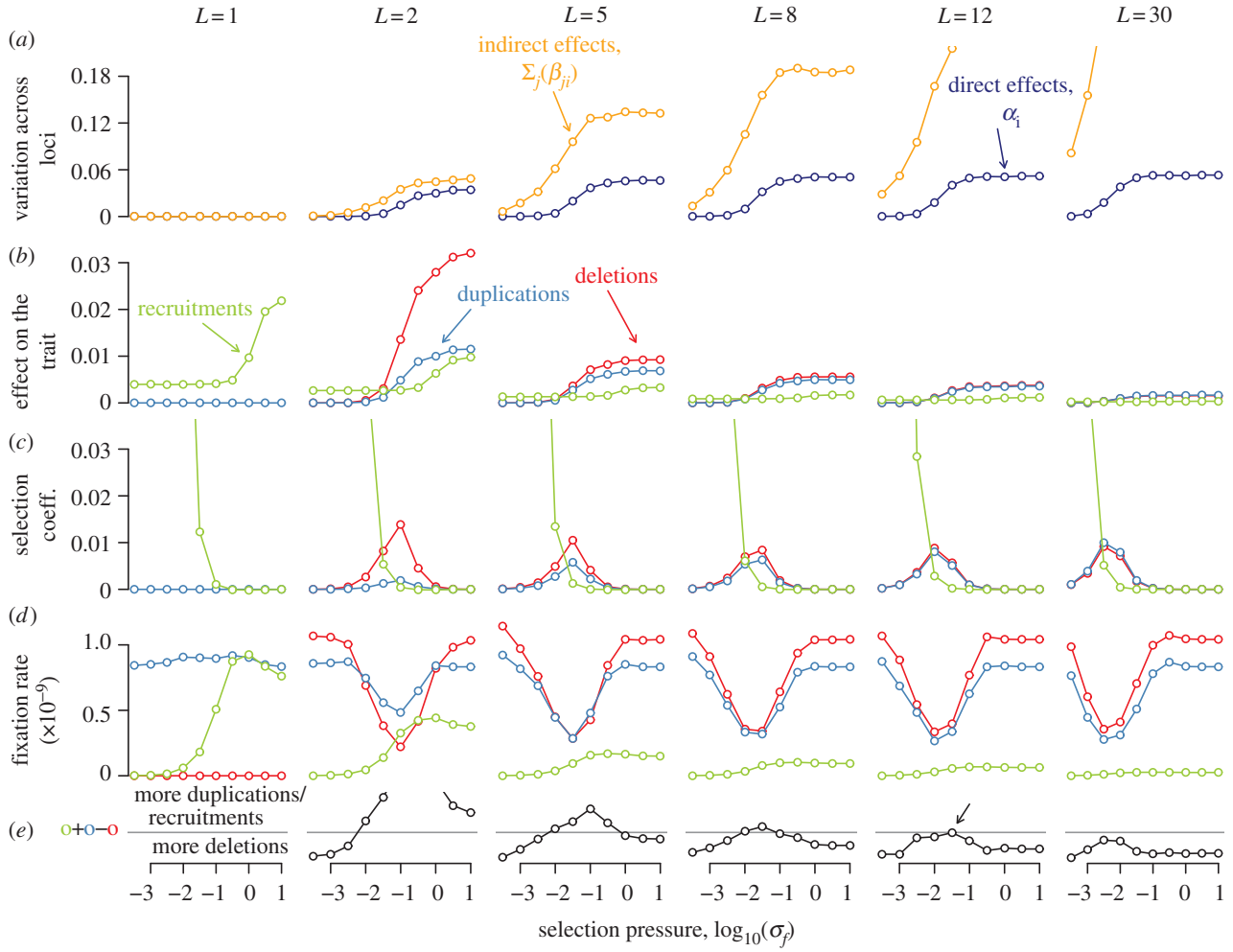

Figure 2. The consequences of gene duplications, recruitments and deletions in a population-genetic model. Populations were initially evolved with a fixed number of controlling loci L (*a*), and we then measured the effects of recruitments, deletions and duplications on the trait value (*b*) and on fitness (*c*). From the latter, we calculated the rate at which deletions, recruitment and duplications enter and fix in the population (*d*), and the resulting rate of change in the number of loci contributing to the trait (*e*). (*a*) For $L > 1$, the variation in direct effects (α_i) and indirect effects among controlling loci ($\Sigma_j(\beta_{ji})$) increases as selection on the trait is relaxed. (*b*) As a consequence of this variation among loci, the average change in the trait value following a duplication or a deletion also increases as selection on the trait is relaxed. (*c*) Changes in the trait value are not directly proportional to fitness costs, because the same change in x has milder fitness consequences when selection is weaker (larger σ_f). As a result, the average fitness detriment of duplications and deletions is highest for traits under intermediate selection. (*d*) Consequently, the fixation rates of duplications and deletions are smallest under intermediate selection. (*e*) The equilibrium number of loci controlling a trait under a given strength of selection is determined by that value of L for which duplications and recruitments on one side, and deletions on the other, enter and fix in the population at the same rate. For example, when $\sigma_f = 10^{-1.5}$, these rates are equal when L is close to 12 (black arrow), so that the equilibrium genetic architecture contains ≈12 loci on average (compare electronic supplementary material, figure S3, black arrow). (Online version in colour.)

required for the evolution of large L, nor does it change the shape of its dependence on the strength of selection.

(b) Intuition for the results

There is an intuitive explanation for the non-monotonic relationship between the selection pressure on a trait and the number of loci that control it. For a trait under weak selection (high σ_f), changes in the trait value have little effect on fitness. Thus, even if deletions, recruitments and duplications change the trait value, these changes are nearly neutral (figure 2). As a result, the number of loci controlling the trait evolves to its neutral equilibrium, which is small because deletions are more frequent than duplications and recruitments (see Methods; figure 1; electronic supplementary material S3). On the other hand, when selection on a trait is very strong (low σ_f), few point mutations, and only those with small effects on the trait, will fix in the population. As a result, all loci have similar contributions to the trait value (figure 2*a*), and so duplications or deletions again have little effect on the trait or on fitness (figure 2*b*,*c*). In this case, the equilibrium number of loci is given by the value expected when deletions and duplications, but not recruitments, are neutral (figure 1; electronic supplementary material S3).

Only when selection on a trait is moderate can variation in the contributions across loci accrue and impact the fixation of deletions and duplications (figure 2*d*), by a process called compensation: a slightly deleterious point mutation at one locus, which perturbs the trait value, segregates long enough to be compensated by point mutations at other loci [32–35]. Compensation increases the variance in the contributions across loci (figure 2*a*), as has been observed for many phenotypes in plants and animals [36]. Owing to this variation, both duplications and deletions move the trait value away from the optimum (figure 2*b*), and so they are mildly deleterious on average (figure 2*c*). Nevertheless, there is a bias favouring duplications over

deletions among the few such events that fix. This bias arises because duplications increase the number of loci in the architecture, which attenuates the effect of each locus on the trait (figure 2*b*). Thus, when selection is moderate, duplications and recruitments fix more often than deletions, driving the number of contributing loci above its neutral expectation (figure 2*d*,*e*). As the number of loci increases the bias is reduced (figure 2*d*,*e*), and so L equilibrates at a predictable value (figure 1; electronic supplementary material S3).

Duplications and recruitments might also be slightly favoured over deletions under intermediate selection, because architectures with more loci also exhibit, in our model, attenuated mutational effects. This effect, which could positively select for an increase in gene copy number [37], is likely weak in our model, because duplications and recruitments are both deleterious on average under intermediate selection, only less so than deletions (figure 2*d*,*e*).

(c) Robustness of results to model assumptions

The predictions of our model—notably, that the number of loci in a genetic architecture and the variance of their allelic contributions are greatest for traits under intermediate selection—are robust to choices of population-genetic parameters. The non-monotonic relation between selection pressure on a trait and the size of its genetic architecture, L, holds regardless of population size, but the location of maximum L is shifted towards weaker selection in larger populations (see electronic supplementary material, figure S5). This result is compatible with our explanation involving compensatory evolution: selection is more efficient in large populations, and so compensatory evolution occurs at smaller selection coefficients. Likewise, when the mutation rate is smaller the resulting equilibrium number of controlling loci is reduced (see electronic supplementary material, figure S6). This result is again compatible with the explanation of compensatory evolution, which requires frequent mutations. Increasing the rate of deletions relative to duplications also reduces the equilibrium number of loci in the genetic architecture, but our qualitative results are not affected even when r_{del} is twice as large as r_{dup} (see electronic supplementary material, figure S7). Finally, increasing the rate of recruitment r_{rec} (or the genome size) increases the number of loci contributing to all traits except those under very strong selection, as expected from figure 2. Our prediction that traits under intermediate selection are encoded by the richest genetic architectures is insensitive to changes in this parameter, and it holds even in the absence of recruitment (see electronic supplementary material, figure S8).

Our analysis has relied on several quantitative-genetic assumptions, which can be relaxed. First, we assumed that all effects of locus i (i.e. α_i and all β_{ij} and β_{ji}) are simultaneously perturbed by a point mutation. Relaxing this assumption so that a subset of the effects are perturbed does not change our results qualitatively (see electronic supplementary material, figure S9). Second, we assumed that point mutations have unbounded effects so that variation across loci can increase indefinitely. To relax this assumption, we made mutations less perturbative to loci with large effects (see Methods). Even a strong mutation bias of this type led to very small changes in the equilibrium behaviour (see electronic supplementary material, figure S10). Third, we assumed no metabolic cost of additional loci, even though additional genes in *Saccharomyces cerevisiae* are known to decrease fitness slightly [38,39]. Nonetheless, including a metabolic cost proportional to L does not alter our qualitative predictions (see electronic supplementary material, figure S11).

We defined the trait value as the average of the contributions α_i across loci (equation (2.1)), as opposed to their sum. This definition reflects the intuitive notion that a gene product's contribution to a trait will generally depend on its abundance relative to all other contributing gene products. If a gene is duplicated or deleted, then the concentration of its products will change relative to other genes contributing to the trait. If the number of contributing genes is already large, this change will have generally have a smaller impact on the phenotype. This attenuating effect of additional contributing loci is also supported by direct empirical data: changing a gene's copy number is known to have milder phenotypic effects when the gene has many duplicates [40,41]. Of course, many alternative definitions of the trait value can include such an attenuating effect, to a greater or lesser extent. We have explored alternative definitions of the trait value, spanning from the average to the sum of contributions across loci, and we find that they generically exhibit the same qualitative results as those obtained under the 'average' definition in equation (2.1) (see electronic supplementary material S1 and figure S12).

Finally, although robust to model formulation and parameter values, our results do depend in part on initial conditions. When selection is strong, the initial genetic architecture can affect the evolutionary dynamics of the number of loci (see electronic supplementary material, figure S14). This occurs because the initial architecture may set dependencies among loci that prevent a reduction of their number. This result indicates that only those architectures of traits under very strong selection should depend on historical contingencies. We have also studied a multi-trait version of our model, where genes participating in other traits can be recruited or lost through mutation. Even though this model features pleiotropy, and the effects of recruitment mutations evolve neutrally instead of being sampled from an arbitrary distribution, our qualitative results remained unaffected (see electronic supplementary material S3 and figure S15).

(d) The dynamics of copy number

Previous models related to genetic architecture have been used to study the evolutionary fate of gene duplicates. These models typically assume that a gene has several subfunctions, which can be gained (neo-functionalization [42]) or lost (subfunctionalization [43,44]) in one of two copies of a gene. Such 'fate-determining mutations' [45] stabilize the two copies, as they make subsequent deletions deleterious. Such models complement our approach, by providing insights into the evolution of discrete, as opposed to continuous or quantitative, phenotypes. Yet there are several qualitative differences between our analysis and previous studies of gene duplication. Most important of these is that our model considers the dynamics of both duplications and deletions, in the presence of point mutations that perturb the contributions of loci to a trait. This coincidence of time scales is important in the light of empirical data [27–30] showing that changes in copy numbers occur at similar rates to point mutations (see electronic supplementary material, table S1). Under these circumstances, a gene may be deleted or acquire a loss-of-function mutation

before a new function is gained or lost. Our model includes these realistic rates, and accordingly we find that duplicates are very rarely stabilized by subsequent point mutations. Instead, the number of loci in a genetic architecture may increase in our model, because compensatory point mutations introduce a bias towards the fixation of duplications as opposed to deletions.

(e) Comparison with empirical data

Like most evolutionary models, our analysis greatly simplifies the mechanistic details of how specific traits influence fitness in specific organisms. As a result, our analysis explains only the broadest, qualitative features of how genetic architectures vary among phenotypic traits, leaving a large amount of variation unexplained. This remaining variation may be partly random (as predicted by the distributions of the number of evolving loci; figure 1), and partly due to ecological and developmental details that our model neglects.

A quantitative comparison between our model and empirical data requires information about the genetic architectures for hundreds of traits (see below for our analysis of expression QTLs). Nevertheless, the qualitative, non-monotonic predictions of our model (figure 1) may help to explain some well-known trends in the genetics of human traits. For instance, in accordance with our predictions, human traits under presumably moderate selection, such as stature (Mendelian inheritance in man, MIM no. 606255 [13,46]), or susceptibility to mid-life diseases such as diabetes (e.g. MIM no. 125853), cancer (e.g. MIM nos. 114480, 176807) and heart disease, are typically complex and highly polygenic, whereas traits under very strong selection, such as those (e.g. mucus composition or blood clotting) affected by childhood-lethal diseases such as cystic fibrosis (MIM no. 219700) or haemophilias (e.g. MIM nos. 306900, 306900) often rely on simple architectures with one or a few loci; and so too traits under very weak selection, such as handedness [47], cerumen moisture (MIM no. 117800 [48]) or bitter taste [49], typically rely on simple architectures. Our analysis provides an evolutionary explanation for these differences, and it delineates the selective conditions under which we expect a Mendelian, as opposed to Fisherian, architecture.

It is important to recognize that our analysis describes the emergence of genetic architectures over evolutionary time scales, which pertain to divergences among species. This perspective complements theories over short time scales, which pertain to segregating variation within a single population, such as the *common disease/common variants* hypothesis, which explains allelic diversity at a human locus in terms of recent population expansion [50,51].

We tested our model for the evolution of genetic architectures by comparison with empirical data on a large number of traits. Such a comparison must, of course, account for the fact that our model describes the true genetic architecture underlying a trait, whereas any QTL study has limited power and describes only the associations detected from polymorphisms segregating in a particular sample of individuals. Accounting for this discrepancy (see below), we compared our model with data from the study of Brem *et al.* [15], who measured mRNA expression levels and genetic markers in 112 recombinant strains produced from two divergent lines of *S. cerevisiae*. For each yeast transcript, we computed the number of non-contiguous markers associated with transcript level, at a given false discovery rate (FDR; see Methods). We also calculated the codon adaptation index (CAI) of each transcript, an index that correlates positively with a gene's expression level [52] and that reflects the environmental conditions encountered during the organism's evolutionary history [53]. CAI correlates with gene dispensability [54,55], and thus it is a decent proxy for the strength of selection on a gene's expression level. We found a striking, non-monotonic relationship between the CAI of a gene and the number of loci linked to variation in its expression (figure 3*a*). Therefore, our analysis of Brem *et al.*'s [15] data indicates that transcript levels under intermediate selection are regulated by more loci than transcripts levels under weak or strong selection.

In order to compare the predictions of our evolutionary model with the empirical data on yeast eQTLs (figure 3*a*), we required to mimic the experimental design of the yeast study. To do so, we first evolved genetic architectures for traits under various amounts of selection (see electronic supplementary material, figure S3), and for each architecture we then simulated a QTL study of the exact same type and power as the yeast eQTL study; that is, we generated 112 crosses from two divergent lines using the empirical yeast genetic map (see electronic supplementary material, figure S2). As expected, the simulated QTL studies based on these 112 segregants detected many fewer loci linked to each trait than in fact contribute to the trait in the true, underlying genetic architecture (figure 3*b* versus figure 1). This result is consistent with previous interpretations of empirical eQTL studies [16].

We can compare the yeast eQTL data only with the qualitative trends in our model, because we lack absolute estimates of selection pressures on yeast transcript abundances. It is likely that the range of simulated selection strengths in figure 3*b* is larger than the range of selection pressures in yeast. For instance, point mutations in genes contributing to traits simulated with $\log_{10}(\sigma_f) \leq -2.5$ have mean selection coefficients $Ns \leq -140$ (see electronic supplementary material, figure S2), which makes them very unlikely to fix (mean fixation probability $\leq 3.5 \times 10^{-5}$), whereas for simulated traits with $\log_{10}(\sigma_f) \geq 0$, point mutations have a mean selection coefficient $Ns \geq -4.5 \times 10^{-3}$ and nearly neutral mean fixation probability ≈ 0.001. This wide range is guaranteed to include biologically relevant selection pressures, but may be broader than the range of selection pressures on yeast transcript levels.

The simulated QTL studies (figure 3*b*) also revealed an important bias: the probability to detect a locus contributing to the architecture of a trait depends on the allelic diversity at that locus, such that the architecture of a trait under weak selection is more likely to be correctly identified in a QTL study than the architecture of a trait under strong selection (see electronic supplementary material, figure S16). Furthermore, our simulations demonstrate that the number of associations detected in such a QTL study also depends on the divergence time between the parental strains used to generate recombinant lines (see electronic supplementary material, figure S17).

Genes with low CAI are typically expressed at low levels and may therefore be prone to more measurement noise. To study this effect, we also simulated measurement noise correlated with selection pressure (see electronic supplementary material, figure S18). Despite these detection biases, the relationship between the selection pressure on a trait and the number of *detected* QTLs in our model (figure 3*b*; electronic supplementary material, figures S18 and S19)

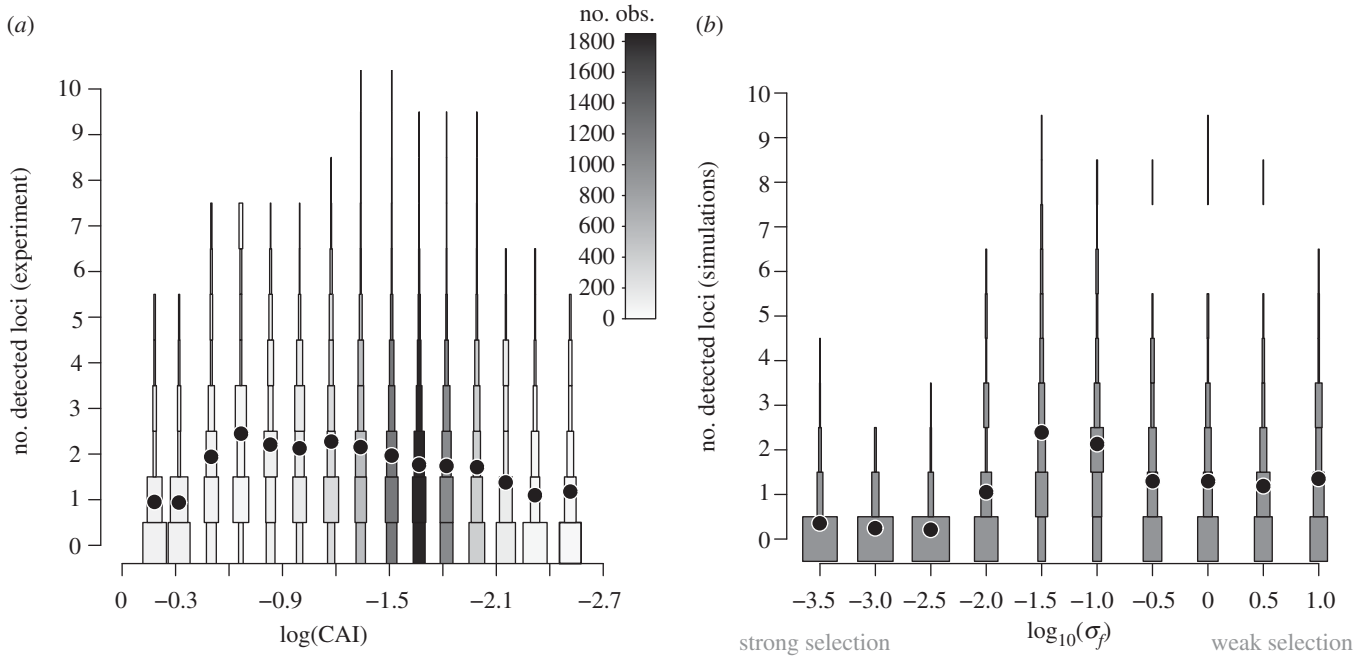

Figure 3. The number of genetic loci controlling a trait inferred (*a*) from real *S. cerevisiae* populations and (*b*) from simulated populations has a non-monotonic relationship with the strength of selection on the trait. (*a*) In the yeast data of Brem *et al.* [15], the largest number of eQTLs were detected for those transcripts (i.e. traits) under intermediate levels of selection (intermediate CAI), whereas fewer eQTLs were detected for transcripts under either weak or strong selection. Transcripts were binned according to their log (CAI) values. Squares represent the distribution of the number of one-way eQTLs identified from the study of Brem *et al*. [15], for traits within each bin of CAI. Grey scale indicates the number of transcripts in each bin. Mean numbers of detected eQTLs are represented by circles. (*b*) For the simulated experiment, we evolved 100 populations of genetic architectures, using the parameters corresponding to the electronic supplementary material, figure S3. From each such population, we then evolved two lines independently for 25 000 generations in the absence of deletions, duplications and recruitment, to mimic the divergent strains used in the yeast cross in [15]. From these two divergent genotypes, we then created 112 recombinant lines following the genetic map of Brem *et al.* [15]. We then analysed the resulting simulated data with *R/qtl* in the same way as we had analysed the yeast data (see electronic supplementary material S2). The distribution of QTLs detected and their means are represented as in figure 1, for each value of selection strength σ_f.

agrees with the relationship observed in the yeast eQTL data (figure 3*a*). Importantly, both of these relationships exhibit the same qualitative trend: traits under intermediate selection are encoded by the richest genetic architectures.

In principle, a non-monotonic pattern similar to the one shown in figure 3*a* could also emerge as a result of the two detection biases that we have described. First, the lack of variation at contributing loci could prevent their detection for traits under very strong selection, whereas elevated measurement noise in weakly expressed genes might also impede the detection of any genetic association. We have quantified these biases in genetic architectures with a fixed number of contributing loci across traits (see electronic supplementary material, figure S20). Our analysis indicates that the non-monotonic pattern observed in the empirical data (figure 3*a*) is unlikely to result from these biases alone. Therefore, detection biases can only accentuate an already non-monotonic pattern in the true architecture.

3. Conclusion

Many interesting developments lie ahead. Here, we have considered only traits under stabilizing selection, whereas some traits might experience fluctuating target phenotypes over the time scales considered, so that periods of stabilizing selection alternate with periods of adaptation. A proper study of the evolution of genetic architectures in fluctuating environments should disentangle these complex dynamics, which is clearly beyond the reach of this paper. Our model is also far too simple to account for tissue- and time-specific gene expression, context-dependent effects, and so on [5,56]. Moreover, we considered the evolution of only haploid genetic architectures. It will be interesting to investigate how architectures evolve when including, for instance, patterns of recombination or dominance and sex-specific effects.

Nonetheless, our work provides intuition that can guide the interpretation and design of empirical studies. Our main finding is that the evolution of genetic architectures is governed by an interaction between the divergence of allelic effects and the number of loci, and this conclusion should hold even in the presence of frequent recombination. More generally, our analysis shows that it is possible to study the evolution of genetic architectures from first principles, to form *a priori* expectations for the architectures underlying different traits, and to reconcile these theories with the expanding body of QTL studies on molecular, cellular and organismal phenotypes.

4. Methods

(a) Model

We described the evolution of genetic architectures using the Wright–Fisher model of a replicating population of size N, in which haploid individuals are chosen to reproduce each generation according to their relative fitnesses. The fitness of an individual with L loci encoding trait value x is $\omega_k = G(x, 0, \sigma_f) \times (1 - L \times c)$, where G denotes the density at x of a Gaussian distribution with mean zero and standard deviation σ_f, and the

second term denotes the metabolic cost of harbouring L loci, which depends on a parameter c. The trait value of such an individual, given the direct contributions α_i and epistatic terms β_{ji}, is described by equation (2.1), where $f_\beta(y) = 2(1 + e^{-s_\beta y})$ is a sigmoidal curve, so that the epistatic interactions either diminish or augment the direct contribution of locus i depending on whether $\sum_j \beta_{ji}$ is positive or negative (see electronic supplementary material, figure S4). In general, loci do not influence themselves ($\beta_{ii} \equiv 0$) and, in the model without epistasis, all $\beta_{ji} \equiv 0$ and $f_\beta \equiv 1$. If an individual chosen to reproduce experiences a duplication at locus i then the new duplicate, labelled k, inherits its direct effect ($\alpha_k = \alpha_i$) and all interaction terms ($\beta_{kj} = \beta_{ij}$ and $\beta_{jk} = \beta_{ji}$ for all $j \neq i,k$), with the interaction terms β_{ik} and β_{ki} initially set to zero. Recruitment occurs with probability r_{rec} per mutation of one of the 6000 genes not contributing to the trait. The initial direct contribution α_i of recruited locus i is drawn from a normal distribution with mean zero and standard deviation σ_m; its interaction terms with other loci (k), β_{ik} and β_{ki}, are initially set to zero. Note that this assumption is relaxed in the multi-locus version of our model, where the direct and indirect effects of recruitments evolve neutrally (see electronic supplementary material, text S3 and figure S15).

In general, a point mutation at locus i changes its contribution to the trait, α_i, and all its epistatic interactions, β_{ij} and β_{ji}, each by an independent amount drawn from a normal distribution with mean zero and standard deviation σ_m. In this model, the impact of a mutation on fitness depends on the ratio σ_f/σ_m, such that an increase in the value of σ_m is strictly equivalent to a proportional decrease of σ_f. The normal distribution satisfies the assumptions that small mutations are more frequent than large ones [57,58], and that there is no mutation pressure on the trait [23]. We relaxed the former assumption by drawing mutational effects from a uniform distribution without qualitative changes to our results (see electronic supplementary material, figure S13). We also used a log-uniform distribution for the mutation effects, such that more mutations have very small effects, again with no qualitative change (see electronic supplementary material, figure S13). In order to relax the latter assumption, we included a bias towards smaller mutations in loci with large effects, so that the mean effect of a mutation at locus i now equals $-b_\alpha \times \alpha_i$ and $-b_\beta \times \beta_{ij}$, respectively for α_i and β_{ij} [59]. We also considered a model in which a mutation at locus i affects only a proportion p_{em} of the values α_i, β_{ij} and β_{ji}.

By default, simulations were initialized with $L = 1$ and $\alpha_1 = 0$; alternative initial conditions were also studied, as shown in electronic supplementary material, figure S14. The code for all simulations and figures presented in the paper was deposited in the Dryad repository (http://dx.doi.org/10.5061/dryad.1f217).

(b) Markov chain for neutral changes in copy number

When deletions and duplications are neutral, and recruitments strongly deleterious, the evolution of the number of loci L in the genetic architecture is described by a Markov chain on the positive integers. The probability of a transition from $L = i$ to $L = i + 1$ equals $r_{dup} \times i$, and that of a transition from i to $i - 1$ is $r_{del} \times i$. We disallow transitions to $L = 0$, assuming that some regulation of the trait is required. We obtained the stationary distribution of L by setting the density of d_1 of individuals in stage 1 to 1 and calculating the density d_i of individuals in the following stages as

$$d_i = \frac{r_{dup} \times (i-1)}{r_{del} \times i} d_{i-1}. \tag{4.1}$$

The equilibrium probability of being in state i was calculated as $p_i = d_i / \sum_{i=1}^{\infty} d_i$, and the expected value of L was calculated as $\sum_{i=1}^{\infty} p_i \times i$. With $r_{dup} = 10^{-6}$ and $r_{del} = 1.25 \times 10^{-6}$, we found an equilibrium expected L of 2.485.

When deletions, duplications and recruitments are all neutral, equation (4.1) can be replaced by

$$d_i = \frac{r_{dup} \times (i-1) + 6000 \times \mu \times r_{rec}}{r_{del} \times i} d_{i-1}. \tag{4.2}$$

Equation (4.2) illustrates the fact that the rates of deletions (which include loss of function mutations) and duplication depend on the number of loci in the architecture, whereas the rate of recruitments does not. With $\mu = 3 \times 10^{-6}$ and $r_{rec} = 5 \times 10^{-5}$, we found an equilibrium expected L of 4.705.

(c) Calculation of $\bar{s}$ and $\overline{p_{fix}}$

We first evolved populations to equilibrium with a fixed number of controlling loci L, and we then measured the effects of deletions, duplications or recruitments introduced randomly into the population. We simulated the evolution of the genetic architecture with L fixed in 500 replicate populations, over 8×10^6 generations for deletions and 10×10^6 generations for duplications, reflecting the unequal waiting time before the two kinds of events. We used 10×10^6 generations for recruitment as well, although different durations did not affect our results. For each genotype k in each evolved population, we calculated the fitness $\omega_k(i)$ of mutants with locus i deleted or duplicated. We calculated the corresponding selection coefficients as $s_k(i) = \omega_k(i)/\langle\omega\rangle - 1$, where $\langle\omega\rangle$ denotes mean fitness in the population. We calculated $\bar{s}$ as the mean across loci and genotypes of $s_k(i)$, weighted by the number of individuals with each genotype. We calculated the probability of fixation of a duplication, deletion or recruitment as

$$p_{fix}(s_k(i)) = \frac{1 - e^{-2s_k(i)}}{1 - e^{-2Ns_k(i)}}, \tag{4.3}$$

and obtained the mean p_{fix} using the same method as for $\bar{s}$.

Rates of deletions and duplications fixing were calculated per locus (figure 2) as r_{del} or r_{dup} times p_{fix}. The total probability of a duplication or a deletion entering the population and fixing is, of course, also multiplied by L. However, recruitment rates remain constant as L changes. Therefore, we divided the rate of recruitments by L in figure 2, for comparison with the *per locus* duplication and deletion rates.

(d) Number of loci influencing yeast transcript abundance

We used the *R/qtl* [60,61] package to calculate logarithm of odds (LOD) scores for a set of 1226 observed markers and 3223 uniformly distributed pseudo-markers separated by 2 cM, by Haley–Knott regression. We calculated the LOD significance threshold for a FDR of 0.2 as the corresponding quantile in the distribution of the maximum LOD after 500 permutations (an FDR of 0.01 and a fixed LOD threshold of three produced qualitatively similar results). The number of detected loci linked to the expression of a transcript was calculated as the number of non-consecutive genomic regions with an LOD score above the threshold. We downloaded *S. cerevisiae* coding sequences from the Ensembl database (EF3 release), and calculated CAI values with the *seqinr* [62] package, using codon weights from a set of 134 ribosomal genes.

Acknowledgements. We thank many anonymous referees, G. Wagner and members of the Plotkin laboratory for constructive feedback.

Funding statement. We gratefully acknowledge support from the Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, the Foundational Questions in Evolutionary Biology Fund (RFP-12-16), the US Army Research Office (grant no. W911NF-12-1-0552) and grant no. D12AP00025 from the US Department of the Interior.

Data accessibility. C++ and R code for simulations and figures is accessible from Dryad at http://dx.doi.org/10.5061/dryad.1f217.

References

1. Hansen TF. 2006 The evolution of genetic architecture. *Annu. Rev. Ecol. Evol. Syst.* **37**, 123–157. (doi:10.1146/annurev.ecolsys.37.091305.110224)
2. Kroymann J, Mitchell-Olds T. 2005 Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**, 95–98. (doi:10.1038/nature03480)
3. Carlborg O, Jacobsson L, Ahgren P, Siegel P, Andersson L. 2006 Epistasis and the release of genetic variation during long-term selection. *Nat. Genet.* **38**, 418–420. (doi:10.1038/ng1761)
4. Rockman MV, Kruglyak L. 2006 Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872. (doi:10.1038/nrg1964)
5. Mackay TFC, Stone EA, Ayroles JF. 2009 The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* **10**, 565–577. (doi:10.1038/nrg2612)
6. Jones A, Arnold S, Bürger R. 2004 Evolution and stability of the g-matrix on a landscape with a moving optimum. *Evolution* **58**, 1639–1654.
7. Carter AJR, Hermisson J, Hansen TF. 2005 The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theor. Popul. Biol.* **68**, 179–196. (doi:10.1016/j.tpb.2005.05.002)
8. Azevedo RBR, Lohaus R, Srinivasan S, Dang KK, Burch CL. 2006 Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* **440**, 87–90. (doi:10.1038/nature04488)
9. de Visser JAGM, Elena SF. 2007 The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.* **8**, 139–149. (doi:10.1038/nrg1985)
10. Fierst JL, Hansen TF. 2010 Genetic architecture and postzygotic reproductive isolation: evolution of Bateson–Dobzhansky–Muller incompatibilities in a polygenic model. *Evolution* **64**, 675–693. (doi:10.1111/j.1558-5646.2009.00861.x)
11. Ungerer MC, Halldorsdottir SS, Modliszewski JL, Mackay TFC, Purugganan MD. 2002 Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. *Genetics* **160**, 1133–1151.
12. Flint J, Mackay TFC. 2009 Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* **19**, 723–733. (doi:10.1101/gr.086660.108)
13. Visscher PM. 2008 Sizing up human height variation. *Nat. Genet.* **40**, 489–490. (doi:10.1038/ng0508-489)
14. Manolio TA *et al*. 2009 Finding the missing heritability of complex diseases. *Nature* **461**, 747–753. (doi:10.1038/nature08494)
15. Brem RB, Storey JD, Whittle J, Kruglyak L. 2005 Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703. (doi:10.1038/nature03865)
16. Brem RB, Kruglyak L. 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102**, 1572–1577. (doi:10.1073/pnas.0408709102)
17. Rockman MV, Skrovanek SS, Kruglyak L. 2010 Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372–376. (doi:10.1126/science.1194208)
18. Emilsson V *et al*. 2008 Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428. (doi:10.1038/nature06758)
19. Ehrenreich IM, Bloom J, Torabi N, Wang X, Jia Y, Kruglyak L. 2012 Genetic architecture of highly complex chemical resistance traits across four yeast strains. *PLoS Genet.* **8**, e1002570. (doi:10.1371/journal.pgen.1002570)
20. Orr HA. 2005 The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6**, 119–127. (doi:10.1038/nrg1523)
21. Rockman MV. 2012 The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1–17. (doi:10.1111/j.1558-5646.2011.01486.x)
22. Yang J *et al*. 2010 Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569. (doi:10.1038/ng.608)
23. Lande R. 1976 The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet. Res.* **26**, 221–235. (doi:10.1017/S0016672300016037)
24. Ewens W. 2004 *Mathematical population genetics: theoretical introduction*. New York, NY: Springer.
25. Proulx SR, Phillips PC. 2006 Allelic divergence precedes and promotes gene duplication. *Evolution* **60**, 881–892.
26. Proulx SR. 2012 Multiple routes to subfunctionalization and gene duplicate specialization. *Genetics* **190**, 737–751. (doi:10.1534/genetics.111.135590)
27. Lynch M *et al*. 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl Acad. Sci. USA* **105**, 9272–9277. (doi:10.1073/pnas.0803466105)
28. Watanabe Y, Takahashi A, Itoh M, Takano-Shimizu T. 2009 Molecular spectrum of spontaneous de novo mutations in male and female germline cells of *Drosophila melanogaster*. *Genetics* **181**, 1035–1043. (doi:10.1534/genetics.108.093385)
29. Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011 High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr. Biol.* **21**, 306–310. (doi:10.1016/j.cub.2011.01.026)
30. van Ommen G-JB. 2005 Frequency of new copy number variation in humans. *Nat. Genet.* **37**, 333–334. (doi:10.1038/ng0405-333)
31. Kuo C-H, Ochman H. 2009 Deletional bias across the three domains of life. *Genome Biol. Evol.* **1**, 145–152. (doi:10.1093/gbe/evp016)
32. Rokyta D, Badgett MR, Molineux IJ, Bull JJ. 2002 Experimental genomic evolution: extensive compensation for loss of DNA ligase activity in a virus. *Mol. Biol. Evol.* **19**, 230–238. (doi:10.1093/oxfordjournals.molbev.a004076)
33. Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010 Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464**, 279–283. (doi:10.1038/nature08691)
34. Kimura M. 1985 The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64**, 7–19. (doi:10.1007/BF02923549)
35. Poon A, Otto S. 2000 Compensating for our load of mutations: freezing the meltdown of small populations. *Evolution* **54**, 1467–1479.
36. Rieseberg LH, Archer MA, Wayne RK. 1999 Transgressive segregation, adaptation and speciation. *Heredity* **83**, 363–372. (doi:10.1038/sj.hdy.6886170)
37. Wagner G, Booth G, Bagheri-Cahichian H. 1997 A population genetic theory of canalization. *Evolution* **51**, 329–347. (doi:10.2307/2411105)
38. Wagner A. 2005 Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**, 1365–1374. (doi:10.1093/molbev/msi126)
39. Wagner A. 2007 Energy costs constrain the evolution of gene expression. *J. Exp. Zool. B* **308**, 322–324. (doi:10.1002/jez.b.21152)
40. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003 Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66. (doi:10.1038/nature01226)
41. Conant GC, Wagner A. 2004 Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. R. Soc. Lond. B* **271**, 89–96. (doi:10.1098/rspb.2003.2560)
42. Ohno S. 1970 *Evolution by gene duplication*. London, UK: George Allen & Unwin Ltd.
43. Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
44. Lynch M, Force A. 2000 The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473.
45. Innan H, Kondrashov F. 2010 The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108. (doi:10.1038/nrg2689)
46. Lango Allen H *et al*. 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838. (doi:10.1038/nature09410)
47. Francks C, DeLisi LE, Fisher SE, Laval SH, Rue JE, Stein JF, Monaco AP. 2003 Confirmatory evidence for linkage of relative hand skill to 2p12-q11. *Am. J. Hum. Genet.* **72**, 499–501. (doi:10.1086/367548)
48. Yoshiura K-I *et al*. 2006 A snp in the abcc11 gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324–330. (doi:10.1038/ng1733)
49. Reed DR *et al*. 2010 The perception of quinine taste intensity is associated with common genetic variants in a bitter receptor cluster on chromosome

12. *Hum. Mol. Genet.* **19**, 4278–4285. (doi:10.1093/hmg/ddq324)
50. Reich DE, Lander ES. 2001 On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510. (doi:10.1016/S0168-9525(01)02410-6)
51. Chakravarti A. 1999 Population genetics: making sense out of sequence. *Nat. Genet.* **21**, 56–60. (doi:10.1038/4482)
52. Sharp P, Li W. 1987 The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295. (doi:10.1093/nar/15.3.1281)
53. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005 Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488. (doi:10.1073/pnas.0501761102)
54. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003 Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235. (doi:10.1101/gr.1589103)
55. Drummond DA, Raval A, Wilke CO. 2008 A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337. (doi:10.1093/molbev/msj038)
56. Ala-Korpela M, Kangas AJ, Inouye M. 2011 Genome-wide association studies and systems biology: together at last. *Trends Genet.* **27**, 493–498. (doi:10.1016/j.tig.2011.09.002)
57. Orr HA. 1999 The evolutionary genetics of adaptation: a simulation study. *Genet. Res.* **74**, 207–214. (doi:10.1017/S0016672399004164)
58. Eyre-Walker A, Keightley PD. 2007 The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618. (doi:10.1038/nrg2146)
59. Rajon E, Masel J. 2011 Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl Acad. Sci. USA* **108**, 1082–1087. (doi:10.1073/pnas.1012918108)
60. Broman KW, Wu H, Sen S, Churchill GA. 2003 *R/qtl*: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890. (doi:10.1093/bioinformatics/btg112)
61. R Development Core Team. 2011 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
62. Charif D, Lobry JR. 2007 SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: Molecules, networks, populations* (eds U Bastolla, M Porto, H Roman, M Vendruscolo), pp. 207–232. New York, NY: Springer.