Frequent non-genic adaptation and gene birth through the interplay of functionality and expression in a population model

Somya Mani
Center for Soft and Living Matter, Institute for Basic Science
Ulsan 44919, Republic of Korea
somyamn@gmail.com

Tsvi Tlusty
Center for Soft and Living Matter, Institute for Basic Science, and
Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology (UNIST)
Ulsan 44919, Republic of Korea
tsvitlusty@gmail.com

ABSTRACT

Over evolutionary timescales, genomic loci switch between functional and non-functional states through processes such as pseudogenization and *de novo* gene birth. Here we ask about the likelihood and rate of functionalization of non-functional loci. We simulate an evolutionary model to look at the contributions of mutations and structural variation using biologically reasonable distributions of mutational effects. We find that a wide range of mutational effects are conducive to functionalization, thus indicating the ubiquity of this process. During functionalization, loci transition from a mutation dominated 'learning' phase to a selection dominated adaptation phase. Interestingly, in the special case of *de novo* gene birth, whereby non-functional loci begin to express a functional product, we find that expression level changes lead to rare, extreme jumps in fitness, whereas sustained adaptation is driven by product functionality. Our work supports the idea that the potential for adaptation is spread widely across the genome, and our results offer mechanistic insights into the process of *de novo* gene birth.

Keywords spontaneous mutation · structural variation · distribution of fitness effects · adaptation · *de novo* gene birth

1 Introduction

At a very coarse level, a genome consists of multiple genomic loci which can be non-functional or loci with functions such as genes, gene regulatory loci, sequences maintaining chromosome structure, etc Consortium et al. [2012]. Currently, genome annotation remains a formidable challenge for both prokaryotes Dimonaco et al. [2022] and eukaryotes Salzberg [2019]. Nevertheless, it is reasonable to assume that on an evolutionary timescale, most genomic loci are in flux across functional and non-functional categories. For example, genes can lose their functionality through pseudogenization Albalat and Cañestro [2016]. In the other direction, there are also many examples across eukaryotes of *de novo* gene birth Van Oss and Carvunis [2019]. In this work, we ask about the fate of non-functional genomic loci.

We approach this question using an evolutionary model and explore how the fitness contribution of a non-functional genomic locus might increase over time due to the effects of accumulating mutations. The distribution of mutational fitness effects (DFE) has been experimentally measured in mutation accumulation studies for various organisms Katju and Bergthorsson [2019]. In our model, we sample biologically reasonable DFEs, using recently measured DFE parameters for *Chlamydomonas* Böndel et al. [2019] as reference. Notably, observations in Böndel et al. [2019] indicate that the DFE of specific regions of the *Chlamydomonas* genome, such as exons, introns or intergenic sequences, are

similar to each other and to the DFE of the whole genome. In general, the DFE is known to vary across different regions of the genome Racimo and Schraiber [2014], and across different species Huber et al. [2017]. We accommodate this diversity by sampling a wide range of DFEs.

Now, over a time scale of millions of years, in addition to small mutations (< 50bp), one can also expect large structural variations (from 50bp upto several megabases) to impact the evolution of genomic loci Mérot et al. [2020]. While the rate of structural variation is estimated to be hundreds of times slower than the rate of small mutations, its effect is likely to be much larger Trost et al. [2021]. Of particular importance to our question is the possibility that the entire genomic locus under consideration gets deleted. Therefore, we test in our model whether sustained fitness increase can occur in the face of locus deletion.

Finally, we consider the particular case of *de novo* gene birth. Recent studies report how new genes gain expression Majic and Payne [2020] and functionality Zhang et al. [2015] over time. Measurements from mutational scans of protein encoding genes indicate that the overall fitness contribution of a gene is a combination of the adaptive value of the expression product, and its expression level Shen et al. [2022]. We envisage that equivalently, during the process of *de novo* gene birth, mutational fitness effects can be decomposed into the effect on adaptive value and the effect on expression level. In the model, we use the DFE, together with empirical measurements of mutational effects on expression to extract a scenario of the evolution of the adaptive value of the expression product.

Overall, we find that a wide range of biologically reasonable DFEs allow functionalization of genomic loci, indicating the ubiquity of this process. Moreover, this gain of functionality occurred despite the antagonistic effects of locus deletion, particularly for the *Chlamydomonas* DFE parameters. In the special case of *de novo* gene birth, our model reveals a short-tailed distribution for mutational effects on adaptive value, thus implying that the rare, extreme mutations that are characteristic of DFEs are instead driven solely by mutational effects on expression level. In contrast, we find that mutations in adaptive value are the major drivers for the sustained fitness increase over evolutionary time. Our results can be tested experimentally using high throughput mutational scans on random initial sequences; such experiments stand to offer quantitative insights into the process of *de novo* gene birth.

2 Model of non-functional locus adaptation

We set up a population genetic framework to model well-mixed populations of fixed size N, composed of asexually reproducing haploid individuals. Fitness of an individual represents exponential growth rate, which is equivalent to the quantities considered in experiments that measure DFEs (e.g., Böndel et al. [2019]). In this work, for any individual i, we consider the evolution of the fitness contribution $F(i)$ of a single locus in its genome. We are interested in the probability that the locus persists in the population, and that its fitness contribution increases above some predetermined threshold. In particular, we examine the special case of *de novo* gene birth, where the fitness contribution can be decomposed into two quantities: functionality, or adaptive value of the expression product ($A(i)$), and its expression level ($E(i)$). Our definition of fitness is not tied to any specific function, and we assume that $F(i) = A(i) \times E(i)$.

The locus of interest is non-genic, with initial fitness $F_0(i) = 0$, and an initial expression level $E_0(i) = 0.001$ for all individuals. The initial expression level captures leaky expression of intergenic regions Clark et al. [2011], which is estimated to be 1000-fold smaller than the level of highly expressed genes Hebenstreit et al. [2011].

Generations in the model are non-overlapping, and the population at time-step $t + 1$ is composed entirely of the offspring of individuals in the time-step t (Fig1(A)). Offspring incur mutations at each time-step, which affect the locus fitness ($\Delta F(i)$). In the case of *de novo* gene birth, $\Delta F(i)$ can be decomposed into mutational effects on adaptive value ($\Delta A(i)$) and expression level ($\Delta E(i)$):

$$F_{t+1}(i) = F_t(i) + \Delta F_t(i) \quad (1)$$

$$\text{and, } F_{t+1}(i) = (A_t(i) + \Delta A_t(i)) \cdot (E_t(i) + \Delta E_t(i)) \quad (2)$$

The mutation rate sets the timescale of the model: a single time-step is roughly the time it takes for one mutation to occur in the locus. For a locus of $\sim$100 base pairs, a single model time-step can range between $100\,\text{years}$ to $100\,000\,\text{years}$ for different organisms (Fig1(B), see also Supplementary Information: Table.1). Offspring can also incur structural variations, which in the model leads to the deletion of the locus in that individual. The probability of locus deletion d represents the rate of structural variation relative to mutation rate.

In the model, the probability that an individual leaves an offspring is proportional to the fitness $F(i)$ of the locus (see Supplementary Information: Section.2 for a discussion of the genomic background). Whenever $F(i) \leq -1$, we

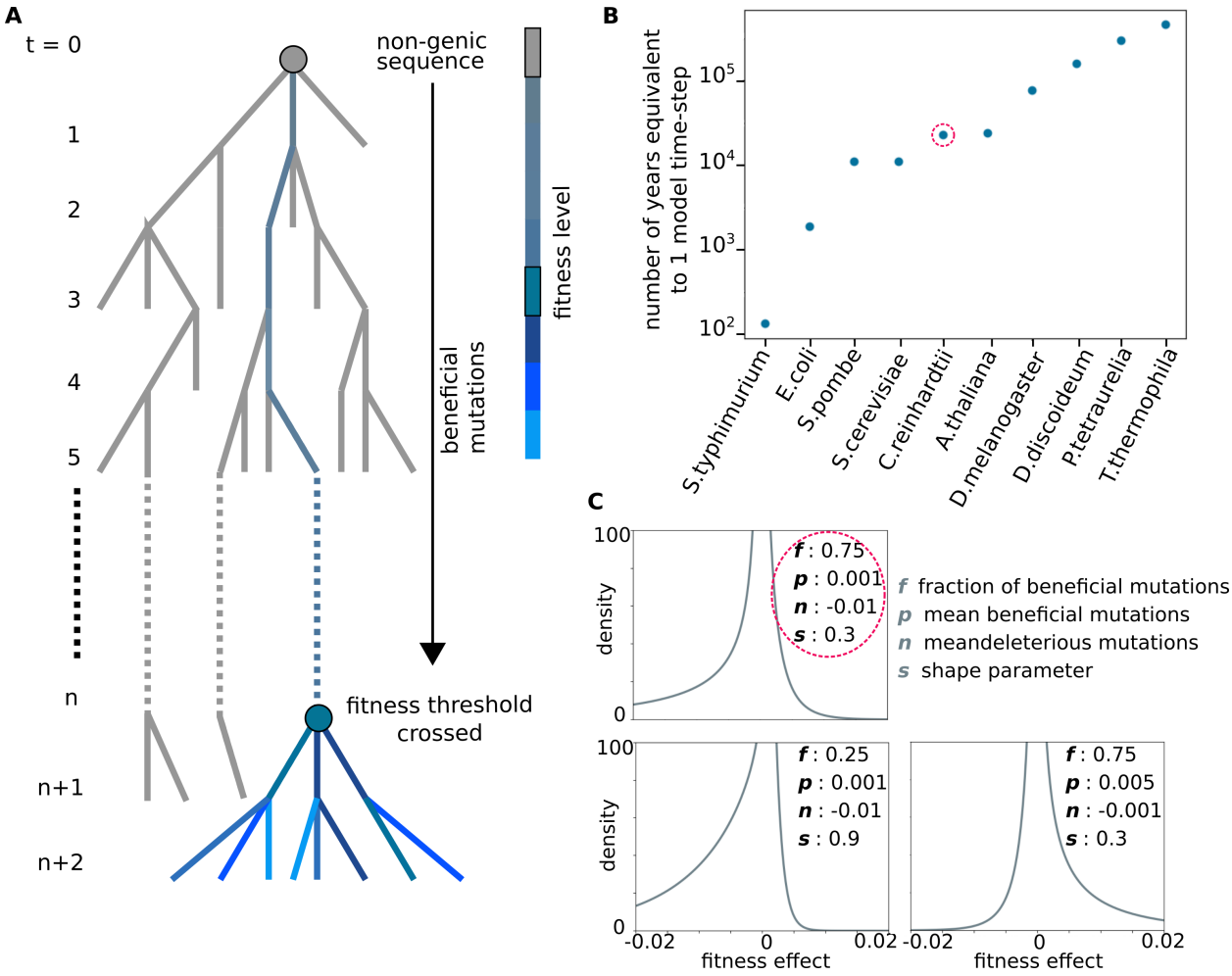

Figure 1: Time-scale and fitness effects of mutations in the model. (A) Phylogenetic tree representing the evolution of a non-genic locus. Time steps t count the generations in the model, which represent the average time for a mutation to occur in the locus. The grey dot at $t = 0$ represents the initial non-genic sequence. Grey branches represent lineages that die out, and colored branches represent the lineage that gets fixed in the population. Fitness levels of colored branches in the fixed lineage are indicated in the color bar. The blue dot at $t = n$ represents the most recent common ancestor of all surviving lineages whose fitness contribution is above the threshold. (B) Estimates of the number of years equivalent to a single time-step of the model in the different species listed on the x-axis. The point representing *Chlamydomonas reinhardtii*, whose DFE is measured in Böndel et al. [2019], is circled in red. See Supplementary Information: Table.1 for calculations. (C) Distribution of fitness effects (DFE) for different values of model parameters (listed for each distribution). The top left panel represents the DFE with parameters closest to those of *C reinhardtii*. The bottom left panel represents the DFE with the most deleterious and least beneficial mutations. The bottom right panel represents the DFE with the most beneficial and least deleterious mutations sampled in this work.

consider the locus lethal and such individuals cannot produce offspring. We update populations for 1000 time-steps, equivalent to 0.1 to 100 million years, depending on the organism and size of the locus (Supplementary Information: Table.1, Method to update population fitness).

Fitness effects of mutations ($\Delta F(i)$) are drawn from the characteristic DFE of the locus (Fig1(C)). Multiple studies indicate that long-tails are important features of DFEs, which can be captured by the general form of long-tailed gamma distributions Eyre-Walker and Keightley [2007]. Therefore, we choose to follow Böndel et al. [2019], and represent DFEs as two-sided gamma distributions, and characterize them using four parameters: (i) average effect of beneficial mutations p, (ii) fraction of beneficial mutations f, (iii) average effect of deleterious mutations n, and (iv) the shape parameter s, where distributions with lower s are more long-tailed. Mutations in the model represent the mutation types included in Böndel et al. [2019], which were single-nucleotide mutations and short indels (insertions or deletions of average length $\leq$ 10 bp) Ness et al. [2015]. Note that here we assume the DFEs of single loci are similar to the DFE across the whole genome, which is the quantity reported in experimental studies. We account for differences in DFEs across species and locations along the genome by sampling across biologically reasonable values of these four parameters p, f, n, s.

We also use empirical measurements to estimate the distribution of mutational effects on expression. Studies indicate that mutational effects on expression from established promoters follow a heavy-tailed distribution Hodgins-Davis et al. [2019]. More relevant to our study of *de novo* gene birth are the recent measurements of mutational effects on expression from *random* sequences Vaishnav et al. [2022], which follow a power law distribution, $\Pr(\Delta E) \sim \Delta E^{-2.25}$. At each time-step, we use the above power law distribution to draw $\Delta E(i)$. We then calculate values of mutational effects $\Delta A(i)$ using equations (1) and (2), given distributions of mutational effects on fitness and on expression level (see Method to update expression level and adaptive value; see also Supplementary Information: FigS.9 for possible deviations from the power-law ΔE distribution due to the very small initial values $E(i)$).

In all, we survey 324 parameter sets – p, f, n, s, the DFE parameters , and d, the probability of locus deletion – (Fig.1(C)). We run 100 replicate populations for each set of parameters for population sizes $N = 100, 1000$ (Surveying the space of DFE and locus deletion parameters in populations of various sizes). At the end of each simulation, we trace the ancestry of each locus in each individual (Tracing ancestry to find fixed mutations) in order to track *fixation events*: a mutant is said to have fixed in the population if the ancestry of all individuals at some time-step t can be traced back to a single individual at some previous time-step $t - t_{\text{fix}}$. During the course of a simulation, populations undergo multiple fixation events. We count the number of replicate populations in which the locus is still retained at time-step $t = 1000$, and the most recent mutant that gets fixed is fitter than a predetermined fitness threshold of 0.1 (Fig1(A)).

Results

Most of the genome is fertile for adaptation

In the absence of locus deletion ($d = 0$), fitness of the last common ancestor crossed the threshold of 0.1 in at least 50% of the replicate populations for a majority (55 out of 81) of DFE parameter sets in $N = 1000$ populations (Fig.2(A), see Supplementary Information: FigS.2 for $N = 100$). The bimodality of the histogram in Fig.2(A) indicates that DFE parameters tend to either be highly conducive, or highly repressive to adaptation. As one can anticipate, the conducive DFE parameter sets tend to have high values for the magnitude (p) and the frequency of beneficial mutations (f), and low values for the magnitude of deleterious mutations (n) and the shape parameter (s) (Fig.2(A),inset and Supplementary Information: FigS.2(A),inset). Particularly, for the *Chlamydomonas* DFE parameters, 97% of the N=1000 replicate populations (52% of N=100 replicate populations) crossed the fitness threshold.

Notably, the four DFE parameters appear to act independently in determining the probability of crossing the fitness threshold. This allows fitness to increase even in populations with small values of parameters f and p, provided the DFE of mutations is long-tailed (i.e., small values of s) (see Supplementary Information: FigS.3). That is, large-effect beneficial mutations are sufficient for adaptive evolution, even when they are rare.

Fitness trajectories involve a transition from a mutation dominated to a selection dominated phase

The fitness trajectories of populations where the fitness threshold is crossed have a typical shape: the population average fitness is initially dominated by the effects of new mutations, which are mostly deleterious, and lead to a decrease in fitness (see Supplementary Information: FigS.4). This is followed by a phase where the effects of selection become visible and average fitness increases roughly linearly. These fitness trajectories are reminiscent of the dynamics of learning through adaptive strategies in gambling problems, where an initial phase of loss of capital due to the cost of learning is followed by recovery Despons et al. [2022].

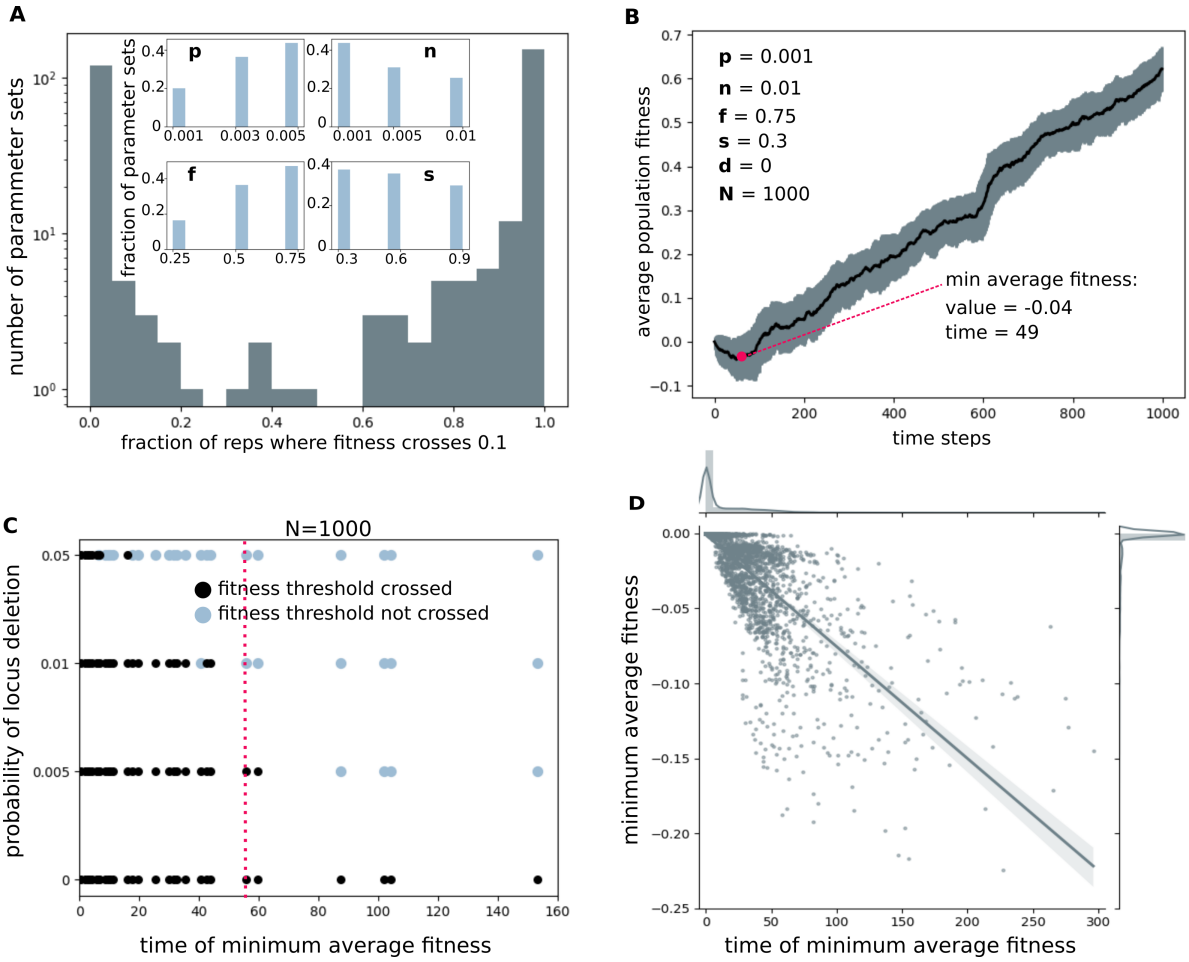

Figure 2: Probability of crossing fitness threshold via accumulating mutations. (A) Histogram for the the fraction of replicate populations that cross the fitness threshold for various parameter sets (p, n, f, s), for $d = 0$ and population size $N = 1000$. Inset: Histograms for the fraction of parameter sets with given values of parameters p, n, f, or s for which more than half of the replicate populations cross the fitness threshold. (B) The trajectory of population average fitness in one of the replicate populations with *Chlamydomonas* DFE parameters (indicated in legend), and no locus deletion ($d = 0$). Average fitness is indicated by the black line, and the grey shading represents standard deviation. The red point indicates the time step at which average fitness has reached minimum value. (C) Scatter plot showing how populations where minimum average fitness is achieved later are more prone to be affected by locus deletion. Each point represents a parameter set. The x-axis indicates the time at which minimum average fitness is achieved (where the averaged is over all the populations with the same DFE parameters, p, n, f, s). Black (blue) points represent parameter sets where $\geq 50\%$ ($< 50\%$) of replicate populations cross the fitness threshold. The dotted red line indicates the *time of minimum average fitness* for *Chlamydomonas* parameters. (D) Scatter plot showing the distribution of minimum average fitness and the time of minimum average fitness for populations which eventually crossed the fitness threshold. The histograms are the marginal distributions of time of minimum fitness and minimum fitness.

Two numbers indicate the point in the trajectory at which selection leads to consistent improvement in fitness: *minimum average fitness* and *time at which minimum fitness is achieved* (Fig. 2(B), Supplementary Information: FigS.5(A)). The DFE parameters, notably p and f are significantly correlated with these quantities (Supplementary Information: Table.2). Moreover, as expected, populations with lower minimum fitness achieve it at later times (Pearson correlation coefficient between *minimum fitness* and *time of minimum fitness* = -0.73) (Fig.2(D)).

Mutations can drive adaptation despite the effect of locus deletion

When $d > 0$, The effect of locus deletion can be understood in terms of a competition between two sub-populations: the sub-population that has lost the locus, and therefore lacks any fitness contribution from it, and the sub-population that retains it (Supplementary Information: FigS.6).

The probability that the sub-population that has lost the locus takes over increase with *time of minimum fitness* as calculated for the case where $d = 0$: the longer the average fitness remains negative, the more probable is the loss of the locus from the whole population. Therefore, fewer replicate populations with DFEs such that minimum fitness is reached later go on to cross the fitness threshold of 0.1 (Fig. 2(C), Supplementary Information: FigS.5(B)). As a consequence, the number of parameter sets for which fitness threshold was crossed in at least 50% of $N = 1000$ replicate populations reduces from 55 at $d = 0$ to 51 at the plausible value of $d = 0.005$, and to 48 and 34 at the inordinately high values of $d = 0.01$ and 0.05, respectively. Particularly, for the *Chlamydomonas* DFE, for which *minimum fitness* and *time of minimum fitness* averaged across all replicate populations are -0.035 and 55.72 respectively, $> 50\%$ of the replicate populations crossed the threshold for $d = 0.005$ (Fig. 2(C), red dotted line).

Functionality drives sustained adaptation, while expression drives extreme mutational events

Our decomposition of fitness into expression level and adaptive value yielded short-tailed exponentially distributed mutational effects on adaptive value (Fig.3(A), Supplementary Information: FigS.7). This indicates that most mutations have little effect on functionality, and mutations with large are extremely rare.

We also looked at correlations between the population averaged fitness trajectories and the average trajectories of expression level and adaptive value. These correlations indicate the contributions of expression level and adaptive value towards the increase of fitness over evolutionary time. We find that in most cases where fitness crosses the 0.1 threshold, the increase in fitness was driven more by the adaptive value than by expression level: the distribution of Pearson's correlation coefficients for adaptive value is sharply peaked at 1, whereas that of expression level is spread broadly (Fig.3(B), Supplementary Information: FigS.8).

As an interesting aside, the empirical measurements that we base our study on do not indicate the level of correlation between the fitness effect and changes in expression level due to mutations; therefore, we proceed with the assumption that $\Delta A(i)$ and $\Delta E(i)$ are independent of each other. In spite of this, over evolutionary time, selection and heritability effectively link fitness and expression level, and impose correlations between their evolutionary trajectories (Supplementary Information: FigS.9).

Overall, we find that sustained adaptation during gene birth is driven more by the product's adaptive value rather than its expression level. At the same time, the extreme mutational effects on fitness, which underlie the long-tails of DFEs, are not driven by changes in the adaptive value of the product, and are instead likely to be entirely driven by changes in expression level. As noted earlier, extreme mutational events become important in facilitating adaptation in cases where beneficial mutations are small and infrequent on average (i.e. small f and p).

Discussion

A majority of studies in genomics and genetics are concerned with the function and evolutionary course of known genes and their regulation. Recent discoveries have attracted focus towards the evolution of non-genic loci; particularly, experimental studies that demonstrate the adaptive potential of random sequences Hayashi et al. [2003], Yona et al. [2018], Lagator et al. [2022]. Furthermore, genomics studies that indicate the frequent occurrence of *de novo* gene birth demonstrate a need for general, theoretical investigations of the evolution of non-genic loci Tautz and Domazet-Lošo [2011].

In this work, we attempt to describe the process of functionalization of non-genic genomic loci in a simple population genetic model. We make use of experimentally measured effects of spontaneous mutations in order to obtain biologically reasonable estimates for the frequency of locus functionalization.

Our model suggests that a wide range of parameters that govern mutational fitness effects (DFE) are conducive to locus functionalization. We find this to be the case despite the antagonistic effects of structural variation that leads to locus deletion. Although the extent of diversity of DFEs across genomic loci and different organisms is not well-known, the range of DFEs surveyed in this work indicates that large swathes of the genome are conducive to adaptation on evolutionary timescales. This result is in line with observations that 80% of the human genome is likely to be functional, while only 3% of the genome contains well-known protein coding exons Consortium et al. [2012]. Our result also supports the proposed prevalence of orphan genes born through *de novo* gene birth Vakirlis et al. [2020].

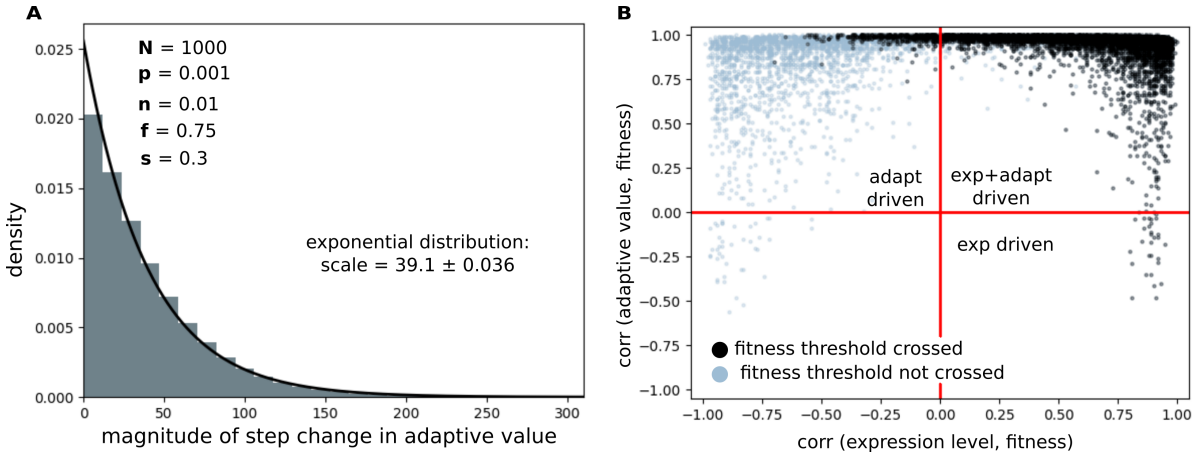

Figure 3: Distribution of $\Delta A(i)$ and the mode of fitness increase. We used the distribution reported in Vaishnav et al. [2022] to generate ΔA in order to obtain trajectories of expression level, e, and adaptive value, a, from which we infer values of ΔA. (A) Histogram of 10^8 ΔA values pooled across all individuals in all replicate populations with *Chlamydomonas* DFE parameters $[p, n, f, s] = [0.001, -0.01, 0.75, 0.3]$. The black line represents an exponential fit. (B) Scatter plot showing how correlated the population averaged fitness trajectory is with the trajectory of population averaged expression level (x-axis) and with the population averaged adaptive value (y-axis). The black points represent populations that cross the fitness threshold, and blue points indicate populations that do not. Overall, the plot contains 81×100 points for each replicate, across all parameter sets for $N = 1000$ populations. The red lines demarcate regions where fitness change is driven by changes in expression level (bottom right), driven by changes in adaptive value (top left), or by both expression level and adaptive value (top right). As expected, there are no black points that lie in the bottom left region, where both adaptive value and expression level are negatively correlated with the fitness trajectory.

The fitness contribution of a gene is a composite function of various molecular mechanisms, for example the accessibility and affinity of the locus to polymerases, the stability, foldability, and interactions of its expression products, etc. Our study of the adaptive value and expression level of *de novo* genes exemplifies how the fitness effects of mutations can be resolved into contributions from underlying mechanisms. Our result also shows how the process of adaptation can be different for *de novo* genes and established genes: we find that in the case of *de novo* gene birth, the increase in fitness was driven more by the adaptive value than by expression level. This effect is likely to be a special feature of *de novo* gene birth, where initially both adaptive value and expression levels are very low. Whereas in the case of established genes, evolution of expression level is known to play a role in adaptation Fraser [2013], Nourmohammad et al. [2017], Blanc et al. [2021].

We built our model to represent naturally evolving populations, where the timescale varies across different organisms, and is set by their respective mutation rates. We assume here that the genomic background, being much larger, evolves at a much faster rate, allowing selection to be solely based on the fitness contribution of the locus of interest. Alternatively, our model can also be used to represent mutation scan experiments such as in Vaishnav et al. [2022], where the genomic background is kept constant. In this case, the generations in the model represent rounds of experiments involving mutagenesis and artificial selection.

The generality of our results is likely to be limited due to dearth of relevant data. Most importantly, we use experimental measurements of DFE and mutational effects on expression that are taken from different organisms: in different organisms, distinct mechanisms produce mutations, therefore the frequencies of different mutation types and its effects may vary across organisms. Although, the leptokurtic nature of DFEs Eyre-Walker and Keightley [2007] and long tailed nature of mutational effects Hodgins-Davis et al. [2019], Vaishnav et al. [2022] on expression have both been observed in independent studies, measurements performed in the same organism could provide important details, for instance the correlations between the effect of a mutation on expression and on fitness. Secondly, the DFE of loci remain constant in our model, while mutational fitness effects are known to vary over evolutionary time due to various causes, such as change of environment, diminishing returns epistasis, etc Sane et al. [2020], Wünsche et al. [2017], Aggeli et al. [2020]. An extended model that includes a consideration of DFE variability would provide valuable insight into the robustness of our results.

We anticipate that our results can be tested and the shortcomings of our model can be addressed through experiments, especially mutational scans such as those in Vaishnav et al. [2022]: For example, one could design experiments that

monitor the fitness effects of mutations on random sequences which also concomitantly detect expression from these random sequences. Alternatively, the evolution of adaptive value of expression products can be directly examined in experiments where random sequences are placed under constitutive, high expression promoters (such as in Hayashi et al. [2003]); in this case the fitness effects of mutations directly correspond to the adaptive value of the product. These experiments, together with theoretical approaches like ours, provide us with means to test and compare the adaptive potential of non-functional genomic sequences, and the general mechanisms of *de novo* gene birth across various organisms.

Methods

2.1 Surveying the space of DFE and locus deletion parameters in populations of various sizes

We scan across DFEs with $p = [0.001, 0.003, 0.005]$, $f = [0.25, 0.5, 0.75]$, $n = [0.001, 0.005, 0.01]$ and $s = [0.3, 0.6, 0.9]$. We look at locus deletion probabilities $d = [0, 0.001, 0.005, 0.01]$. And we look at populations of sizes $N = [100, 1000]$. For each parameter set, we simulate 100 replicate systems. In all, we look at 64 800 systems. All codes used to generate and analyze data are written in Python3.6.

2.2 Method to update population fitness

For a population of size N, fitness of individuals at time-step t are stored in the vector $F_t \in \mathbb{R}^{N\times 1}$, where the fitness of any individual i is $F_t(i)$. We also keep track of the individuals that have lost the locus due to deletion in the vector $L_t \in [0,1]^{N\times 1}$, such that $L_t(i) = 1$ implies that individual i contains the locus at time-step t, and $L_t(i) = 0$ implies individual i has lost the locus. Note that $L_t(i) = 0$ automatically implies $F_t(i) = 0$.

In the model, only individuals with fitness > -1 are viable, and capable of producing progeny. And individuals in the current population that produce progeny are chosen on the basis of their relative fitness. Let minfit_t be the minimum fitness among viable individuals in F_t.

We define $\text{allfit}_t = \sum_j (1 + F_t(j) - \text{minfit}_t)$, for j such that $F_t(j) > -1$. The normalized relative fitness of individuals is then given by $\text{relfit}_t \in [0,1]^{NX1}$, where

$$\text{relfit}_t(i) = \frac{1 + F_t(i) - \text{minfit}_t}{\text{allfit}_t}, \qquad \forall i \text{ s.t. } F_t(i) > -1$$
$$\text{and, } \text{relfit}_t(i) = 0, \qquad \forall i \text{ s.t. } F_t(i) \leq -1$$

Therefore, even if $F_t(i) = 0$, $\text{relfit}_t(i)$ can be non-zero if $\text{minfit}_t < 0$.

Let $\text{Anc}_{t+1} \in \mathbb{N}^{NX1}$ be the list of individuals chosen from the current time-step t to leave progeny. In other words, Anc_{t+1} is the list of ancestors of the population at time-step $t+1$. Here, $Pr(\text{Anc}_{t+1}(j) = i) \propto \text{relfit}_t(i), \forall i, j \leq N$.

Progeny of the current population incur mutations. The mutation effects are drawn from 2-sided gamma distributions governed by the parameters *p* (average effect of beneficial mutations), *f* (fraction of beneficial mutations), *n* (average effect of deleterious mutations), and *s* (shape parameter). The values of fitness effects of mutations incurred by each individual at time-step t is stored in $\text{mut}_t \in \mathbb{R}^{N\times 1}$, where

$$\text{mut}_t(i) = \Gamma\left(s, \frac{p}{s}\right) \iff \text{Ber}(f) = 1\,,$$
$$\text{and, } \text{mut}_t(i) = \Gamma\left(s, \frac{n}{s}\right) \iff \text{Ber}(f) = 0\,.$$

Here $\Gamma(\kappa, \theta)$ represents a number drawn from the gamma distribution with shape parameter κ and scale parameter θ, and $\text{Ber}(f)$ is the Bernoulli random variable which equals 1 with probability f.

Progeny can also lose the locus with probability d. Thus, the updated fitness levels of the population is given by $F_{t+1}(i) = 0$, if $Anc_{t+1}(i)$ did not contain the locus, or if the individual loses the locus in the current time step. Otherwise, $F_{t+1}(i) = F_t(\text{Anc}_{t+1}(i)) + \text{mut}_t(i)$.

2.3 Method to update expression level and adaptive value

In the model, we assume $F(i) = A(i) * E(i)$ for any individual i. For a population of size N, expression levels of the locus at time-step t are stored the vector $E_t \in \mathbb{R}^{N\times 1}$, where the expression level of some individual i is $E_t(i)$. For an individual that has lost the locus due to deletion, $L_t(i) = 0$, which automatically implies $E_t(i) = 0$.

Initially, the expression level of the locus across the population is distributed around 0.001, and reflects leaky expression. At each time step, the expression levels across the population change as individuals are selected and their progeny incur mutations.

The effect of mutations on expression level incurred by each individual at time-step t is stored in $\Delta E_t \in \mathbb{R}^{N\times 1}$. The magnitude of $\Delta E_t(i)$ are drawn from a power law distribution such that $Pr(|\Delta E_t(i)| = x) = x^{-}2.25$ for $x \geq 0$. We assume that a $\Delta E_t(i)$ is negative with probability 0.5.

The updated expression levels of the population are therefore given by $E_{t+1}(i) = 0$, if $Anc_{t+1}(i)$ did not contain the locus, or if the individual loses the locus in the current time step. If the individual does contain the locus, $E_{t+1}(i) = E_t\left(\mathrm{Anc}_{t+1}(i)\right) + \Delta E_t(i)$.

Note that the values of expression level in the model are bounded within $[0.001, 1]$ corresponding to leaky expression and maximal possible expression respectively. In the simulation, whenever $E_{t+1}(i) < 0.001$ or $E_{t+1}(i) > 1$, we reset it to 0.001 and 1, respectively. Since the initial expression levels are very low, $E_{t+1}(i)$ never crossed 1 in any simulation. In a run of 1000 time steps, $E_{t+1}(i)$ crosses 0.001 on average 40 times (Supplementary Information: FigS.10).

We then calculate the corresponding changes in the adaptive value of the locus at each time step: $A_t(i) = F_t(i)/E_t(i)$. From this, we can calculate the change in adaptive value due to mutation as $\Delta A_t(i) = A_{t+1}(i) - A_t\left(Anc_{t+1}(i)\right)$.

2.4 Tracing ancestry to find fixed mutations

In order to find the fitness value of the mutant fixed in the population at time-step t, we start with the list of ancestors of individuals Anc_t at time-step t.
Let $X_t = \{i, \forall i \in \mathrm{Anc}_t\}$ be the set of unique ancestor identities. We then recursively find $X_{t-n} = \{i, \forall i \in \{\mathrm{Anc}_{t-n}(j),\ \forall j \in X_{t-n+1}\}\}$ as the set of unique ancestor identities for $n = 1, 2, 3...t_0$, where X_{t-t_0} is the first singleton set encountered. This set contains a single individual at time-step $t - t0 - 1$, whose mutations are inherited by every individual at time-step t. And the fitness value of the mutant fixed in the population at time-step t is then $F_{t-t_0-1}(i),$ where $i \in X_{t-t0}$.

Acknowledgements

We thank John McBride and Luca Peliti for very helpful discussions. This work was funded by the Institute for Basic Science, Grant IBS-R020.

Author contributions

SM: conceived the project, designed research, developed models, performed simulations, analysed data and wrote the manuscript. TT: conceived the project, supervised research, and wrote the manuscript.

Competing interests

The authors declare they have no competing interests.

References

ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489 (7414):57, 2012.

Nicholas J Dimonaco, Wayne Aubrey, Kim Kenobi, Amanda Clare, and Christopher J Creevey. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics*, 38(5):1198–1207, 2022.

Steven L Salzberg. Next-generation genome annotation: we still struggle to get it right. *Genome biology*, 20(1):1–3, 2019.

Ricard Albalat and Cristian Cañestro. Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379–391, 2016.

Stephen Branden Van Oss and Anne-Ruxandra Carvunis. De novo gene birth. *PLoS genetics*, 15(5):e1008160, 2019.

Vaishali Katju and Ulfar Bergthorsson. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biology and Evolution*, 11(1):136–165, 2019.

Katharina B Böndel, Susanne A Kraemer, Toby Samuels, Deirdre McClean, Josianne Lachapelle, Rob W Ness, Nick Colegrave, and Peter D Keightley. Inferring the distribution of fitness effects of spontaneous mutations in chlamydomonas reinhardtii. *PLoS biology*, 17(6):e3000192, 2019.

Fernando Racimo and Joshua G Schraiber. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genet*, 10(11):e1004697, 2014.

Christian D Huber, Bernard Y Kim, Clare D Marsden, and Kirk E Lohmueller. Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences*, 114(17):4465–4470, 2017.

Claire Mérot, Rebekah A Oomen, Anna Tigano, and Maren Wellenreuther. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7):561–572, 2020.

Brett Trost, Livia O Loureiro, and Stephen W Scherer. Discovery of genomic variation across a generation. *Human Molecular Genetics*, 30(R2):R174–R186, 2021.

Paco Majic and Joshua L Payne. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Molecular biology and evolution*, 37(4):1165–1178, 2020.

Wenyu Zhang, Patrick Landback, Andrea R Gschwend, Bairong Shen, and Manyuan Long. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome biology*, 16(1):1–14, 2015.

Xukang Shen, Siliang Song, Chuan Li, and Jianzhi Zhang. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, 2022.

Michael B Clark, Paulo P Amaral, Felix J Schlesinger, Marcel E Dinger, Ryan J Taft, John L Rinn, Chris P Ponting, Peter F Stadler, Kevin V Morris, Antonin Morillon, et al. The reality of pervasive transcription. *PLoS biology*, 9(7): e1000625, 2011.

Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden, and Sarah A Teichmann. Rna sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular systems biology*, 7(1):497, 2011.

Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, 2007.

Rob W Ness, Andrew D Morgan, Radhakrishnan B Vasanthakrishnan, Nick Colegrave, and Peter D Keightley. Extensive de novo mutation rate variation between individuals and across the genome of chlamydomonas reinhardtii. *Genome Research*, 25(11):1739–1749, 2015.

Andrea Hodgins-Davis, Fabien Duveau, Elizabeth A Walker, and Patricia J Wittkopp. Empirical measures of mutational effects define neutral models of regulatory evolution in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 116(42):21085–21093, 2019.

Eeshit Dhaval Vaishnav, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory dna. *Nature*, 603(7901):455–463, 2022.

Armand Despons, David Lacoste, and Luca Peliti. Adaptive strategy in kelly's horse races model. *arXiv preprint arXiv:2201.03387*, 2022.

Yuuki Hayashi, Hiroshi Sakata, Yoshihide Makino, Itaru Urabe, and Tetsuya Yomo. Can an arbitrary sequence evolve towards acquiring a biological function? *Journal of molecular evolution*, 56(2):162–168, 2003.

Avihu H Yona, Eric J Alm, and Jeff Gore. Random sequences rapidly evolve into de novo promoters. *Nature communications*, 9(1):1–10, 2018.

Mato Lagator, Srdjan Sarikas, Magdalena Steinrueck, David Toledo-Aparicio, Jonathan P Bollback, Calin C Guet, and Gašper Tkačik. Predicting bacterial promoter function and evolution from random sequences. *Elife*, 11, 2022.

Diethard Tautz and Tomislav Domazet-Lošo. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12 (10):692–702, 2011.

Nikolaos Vakirlis, Anne-Ruxandra Carvunis, and Aoife McLysaght. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500, 2020.

Hunter B Fraser. Gene expression drives local adaptation in humans. *Genome research*, 23(7):1089–1096, 2013.

Armita Nourmohammad, Joachim Rambeau, Torsten Held, Viera Kovacova, Johannes Berg, and Michael Lässig. Adaptive evolution of gene expression in drosophila. *Cell reports*, 20(6):1385–1395, 2017.

Jennifer Blanc, Karl AG Kremling, Edward Buckler, and Emily B Josephs. Local adaptation contributes to gene expression divergence in maize. *G3*, 11(2):jkab004, 2021.

Mrudula Sane, Gaurav D Diwan, Bhoomika A Bhat, Lindi M Wahl, and Deepa Agashe. Shifts in mutation spectra enhance access to beneficial mutations. *bioRxiv*, 2020.

Andrea Wünsche, Duy M Dinh, Rebecca S Satterwhite, Carolina Diaz Arenas, Daniel M Stoebel, and Tim F Cooper. Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. *Nature Ecology & Evolution*, 1 (4):1–6, 2017.

Dimitra Aggeli, Yuping Li, and Gavin Sherlock. Changes in the distribution of fitness effects and adaptive mutational spectra following a single first step towards adaptation. *bioRxiv*, 2020.