HYPOTHESIS
Insights & Perspectives

de novo gene birth as an inevitable consequence of adaptive evolution

Somya Mani[1] | **Tsvi Tlusty**[1,2]

[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, Republic of Korea

[2]Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

Correspondence

Email: tsvitlusty@gmail.com

Funding information
Institute for Basic Science, Grant IBS-R020.

Phylostratigraphy suggests that new genes are continually born *de novo* from non-genic sequences, and the genes that persist found new lineages, contributing to the adaptive evolution of organisms. While recent evidence supports the view that *de novo* gene birth is frequent and widespread, the mechanisms underlying this process are yet to be discovered. Here we hypothesize and examine a potential general mechanism of gene birth driven by the accumulation of beneficial mutations at non-genic loci. To demonstrate this possibility, we model this mechanism within the boundaries set by current knowledge on mutation effects. Estimates from this analysis are in line with observations of recurrent and extensive gene birth in genomics studies. Thus, we propose that, rather than being inactive and silent, non-genic regions are likely to be dynamic storehouses of potential genes.

KEYWORDS
de novo gene birth, spontaneous mutation, distribution of fitness effects, adaptation

Introduction

Broadly, a gene is defined as a sequence in the genome which yields phenotypic traits through regulatory interactions of its products with other genes and the environment [1]. For long, the answer to the question of origin of genes was taken for granted; it was believed that a basic set of 'founder genes', numbering some thousands, originated a long time ago, and all new genes are exclusively derived from these founder genes. In contrast to this 'genes come from genes' picture, genomic evidence indicates that 10-30 % of all genes across eukaryotic genomes are orphan genes, for which no ho-

mology can be detected with any established conserved genes [2]. While it is possible that orphan genes are products of duplication and rapid divergence of conserved genes, recent studies indicate that most orphan genes are likely to be generated *de novo* , starting from previously non-genic sequences [3]. Moreover, phylostratigraphy suggests that such new genes are generated continuously [2]. Much of our current understanding of the rates and extent of *de novo* gene birth comes from genomics studies. But to understand the general mechanisms underlying the process, there is a need for independent theoretical models that build upon basic evolutionary processes. In this work, we propose one such basic mechanism — *de novo* gene birth through the accumulation of beneficial mutations— and demonstrate it using a simple mathematical model.

Known *de novo* genes display some intriguing patterns: new genes are born preferably in genomic regions with high GC content and near meiotic hot spots. In animals, new genes are more likely to be expressed in the brain and testis [4]. Interestingly, these cells and genomic regions are also especially prone to pervasive, leaky transcription [5]. These observations point to a possible mechanism of gene birth [6]— non-genic loci, made visible to natural selection by pervasive expression, can be driven to evolve adaptively and gain new functions and thereby lead to *de novo* gene birth [7].

Two simple studies, taken together, lend support to such a mechanism: first, random sequences can gain functionality, provided they are consistently expressed [8]. And second, it was demonstrated that new promoters could easily evolve in *E. coli* [9]. These studies highlight the possibility that non-genic sequences can, in stages, gain the hallmarks of genes: *regulated expression* and *functionality*.

We draw on these observations and further propose that gene birth can be understood as an *inevitable* consequence of adaptive evolution of non-genic sequences. To illustrate this, we present a blueprint for a minimal model of gene birth that uses characteristics of spontaneous mutations, the simplest units of adaptive evolution, as its building blocks. Specifically, we consider the process by which a locus that initially provides very low fitness advantage, corresponding to leaky expression, starts to accumulate beneficial mutations. We conjecture that this process reflects the early stages of *de novo* gene birth.

In the model, we define both genes and mutations in a coarse-grained manner, purely in terms of their fitness contributions. Practically, genes can be defined at many levels; genes are currently described as conserved genomic sequences that produce functional products. Function and regulated expression are fundamental hallmarks of a gene, and gene birth is essentially the process of concerted evolution of these aspects. But the dearth of quantitative data that describe the evolution of expression levels and potential functionality of non-genic sequences constrains us to employing an abstract definition of genes. Nevertheless, this simplification allows us to leverage currently available data to produce biologically reasonable estimates.

Similarly, we describe mutations also in terms of their fitness effects. Experimentally, fitness effects of spontaneous mutations are assessed through mutation accumulation studies. These studies directly [10], or indirectly [11] allow inference of the fitness effects of single mutations, thereby yielding a distribution of fitness effects (DFE). The key assumption of our model is that the DFE of small loci (100-1,000 base pairs) are similar in form to the DFE across the whole genome, which is the quantity measured in mutation accumulation experiments. In particular, we assume that the DFE of loci that start out as non-genic and are evolving adaptively, can be captured using the same parameters that are used to describe the DFE of whole genomes. These assumptions are supported by observations in [11], that the DFE of specific regions of the genome, such as exons, introns or intergenic sequences, are similar to each other and to the DFE of the whole genome. Now in general, the DFE is known to differ across different regions of the genome [12], and across different species [13]. We accommodate this diversity by sampling a wide range of DFEs, which differ in the frequency and size of benefi-

cial and deleterious mutations, and also in the shape of the distribution, to assess the conditions under which spontaneous mutations can lead to gene birth.

Our simple analysis indicates that gene birth should be highly likely for a wide range of DFEs, given a time frame of millions of years. Especially, we find that the presence of rare, large-effect mutations can potentially compensate for beneficial mutations being small and infrequent on average. We also tested the more realistic scenario where the DFE of a genomic locus fluctuates over long periods of time; under these conditions, gene birth becomes virtually inevitable.

Thus, we propose the intriguing hypothesis that *de novo* gene birth through the adaptive evolution of non-genic sequences is practically unavoidable. We also discuss experiments to test the consequences of this hypothesis. We anticipate that in the future, experiments would characterize not only the fitness effects of mutations, but also distinguish between their effect on expression level of loci and functionality of expression products. Such data should inform more detailed models that capture the essence of genes, and therefore of the process of *de novo* gene birth.

An adaptive model of gene birth

We use the Wright-Fisher framework to model well-mixed populations of fixed size N, composed of asexually reproducing haploid individuals. For each individual i, we consider the fitness contribution F_i of a single locus in its genome. Here, fitness represents exponential growth rate, which is equivalent to the quantities considered in experiments that measured DFEs (e.g., [11]). Since our definition of genes is not tied to any specific function, we describe a locus as *genic* if it consistently contributes a fitness advantage above a predetermined *genic threshold* and non-genic otherwise.

Traditionally, the Wright-Fisher model is used to study the evolution of pre-existing genes; here we apply the model to non-genic loci and test the predictions it makes for *de novo* gene birth. Initially, the distribution of F_i is centered around 0 and any small fitness effect is due to leaky expression. In each subsequent time-step $t+1$, the population is composed entirely of the offspring of individuals in the current time-step t (Fig1(A)). The probability that an individual leaves an offspring is proportional to the fitness F_i of the locus, and individuals with $F_i \leq -1$ cannot produce offspring.

Additionally, offspring in the model incur mutations. This sets the timescale of the model according to the mutation rate of the organism: a time-step in the model is roughly the time it takes for a mutation to occur in the locus. For a locus of ~100 base pairs, a single model time-step can range between 100 – 100,000 years for different organisms (Fig1(B), see also FigS3).

The fitness effects of mutations are drawn from the characteristic DFE for the locus (Fig1(C)). Multiple studies indicate that long-tails are important features of DFEs, and the gamma distribution is a general form which can describe such long-tailed distributions [14]. Therefore, we choose to follow [11], 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. Note that, although experimental studies report DFEs across the whole genome, here we assume that the characteristics of the DFE of single loci are similar. The model describes the mutation types included in [11], which were single-nucleotide mutations and short indels (insertions or deletions of average length $\leq$ 10 bp) [15]. We account for differences in DFEs across species and locations on the genome by sampling across biologically reasonable values of these four parameters p, f, n, s (see Surveying the space of DFEs in populations of various sizes). In all, we survey 225 parameter sets, and run 100 replicate populations for each set of DFE parameters.

We update populations for 2,500 time-steps, equivalent to 0.2-200 million years, depending on the organism and size of the locus (see Method to update population fit-

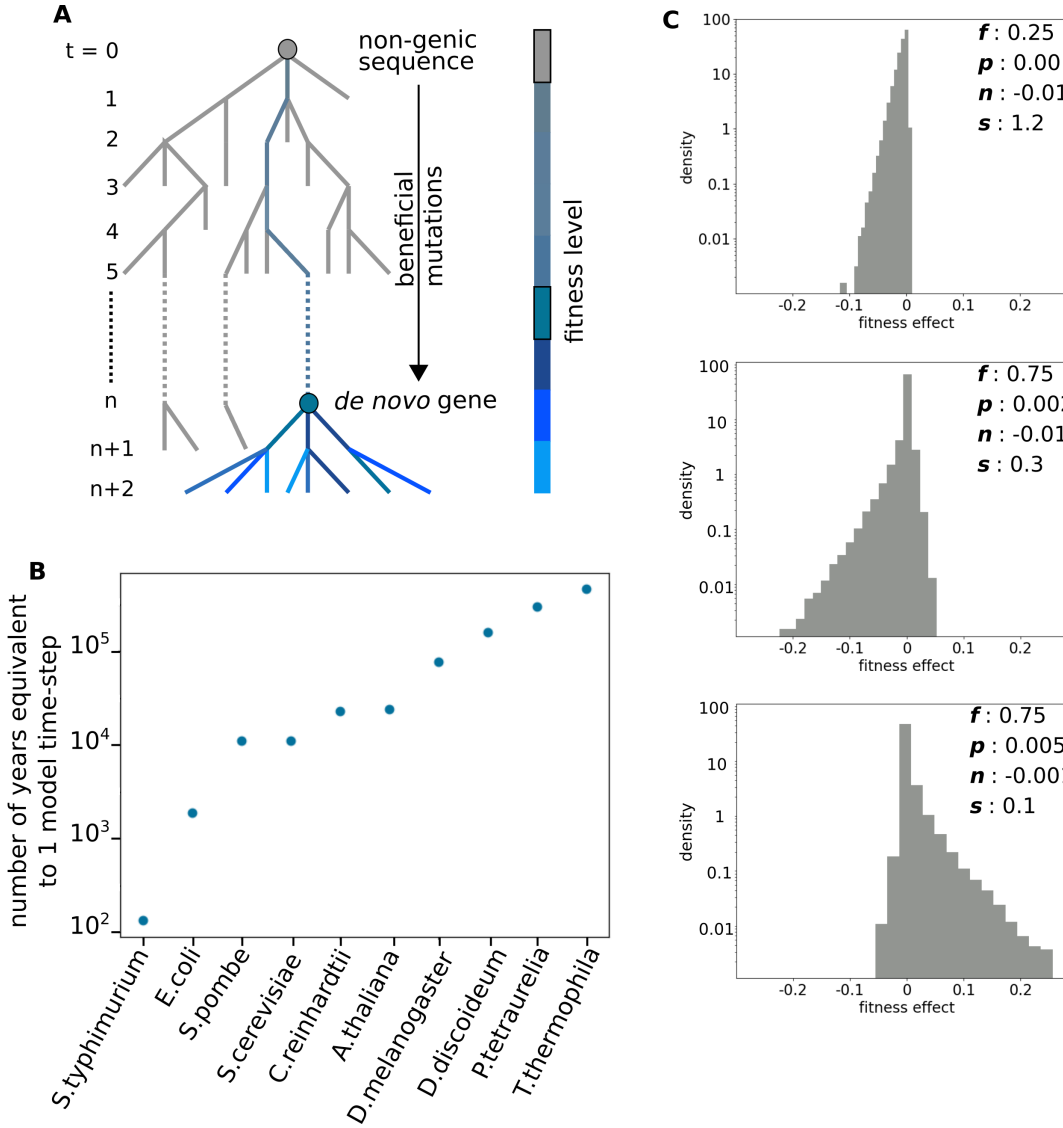


FIGURE 1 Time-scale and fitness effects of mutations in the model. (A) Phylogenetic tree representing the evolution of the non-genic locus into a *de novo* gene. Time steps t count the generations in the model, which represent the average time for a mutation to occur in the locus. The grey dot at $t = 0$ represents the initial non-genic sequence. Grey branches represent lineages that die out, and colored branches represent the lineage that gets fixes in the population. Fitness levels of colored branches in the fixed lineage are indicated in the color bar; the fitness level corresponding to non-genic sequences and to the genic threshold are outlined as the grey and blue box respectively. The blue dot at $t = n$ represents the most recent common ancestor of all surviving lineages whose fitness contribution is above the genic threshold. In the model, gene birth is said to have occurred at time step $t = n$. (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. See FigS1 for calculations. (C) Distribution of fitness effects (DFE) for different model parameters. The top panel represents the DFE with the most deleterious and least beneficial mutations. The bottom panel represents the DFE with the most beneficial and least deleterious mutations sampled in this work. The middle panel represents the DFE with parameters closest to those reported in [11]. Values of parameters used to construct these DFEs are given alongside each histogram.

ness for finite populations). In finite populations, we can trace the ancestry of each locus in each individual (see Tracing ancestry and finding the fitness of the fixed mutation), which allows us 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 say that *de novo* gene birth occurs when the most recent mutant that gets fixed in the population is fitter than the predetermined *genic* threshold (Fig1(A)). For infinite populations, we model the evolution of the locus as a stochastic process (see Method to update population fitness for infinite populations). In this case, we say that gene birth occurs when the average fitness of the population is greater than the *genic* threshold.

We verified that the model dynamics conform qualitatively with known results from the population genetics literature. For example, it is well-known that small populations are generally subject to stronger genetic drift, which makes it harder for fitter mutants to fix ([16], chapter 2). Consistently, in our simulations of populations of size N = 100, 1,000 and 5,000, the probability of fixation of fitter *de novo* mutants that arise in smaller populations is lower (Table.1(row1), FigS2(A), FigS3(A)). Additionally, mathematical models of the Moran process indicate that fitter mutants have a lower fixation probability and a shorter fixation time in smaller populations [17]. In our model, across all fixation events, the mean fitness difference between the current fixed mutation and the previously fixed mutation tends to be much higher in larger populations (FigS2(B), FigS3(B)). And fixation occurs much faster in $N = 100$ populations (Table.1(row2), FigS2(C), FigS3(C)).

In the following, we project these known results to the problem of gene birth, and derive expected bounds for the time scale and frequency of *de novo* gene birth.

TABLE 1 Dynamics of fixation of beneficial mutations in finite populations. (Row1) The fraction of systems in which a fitter mutant fixed in the population at least 75% of the time across 2,500 time-steps. (Row2) The average of mean-fixation time across 2,500 time-steps.

	N=100	N=1000	N=5000
Fraction of systems where a fitter mutant fixed at least 75 % of the time	0.57	0.70	0.77
Mean time-steps to fixation	104.2	226.2	296.4

Predictions from the adaptive model

| Most of the genome is fertile ground

The parameters encoding different DFE represent the variety of different genomic regions across different species. We find that a majority of parameters in our survey are conducive to gene birth (Fig2(A), FigS4): across N = 100, 1,000 populations, 60.9 % and 72.4 %, respectively, of all parameters led to gene birth in all 100 replicate systems. Thus, our model suggests that gene birth due to spontaneous mutations should be a universal process.

Qualitatively, dynamics of gene birth in the model concur with results on fixation dynamics of beneficial mutations (Table.1): gene birth is more prevalent in larger populations (Table.2 (row1)), and all parameters that allow gene birth in a small population also allow it in larger populations. Also, in finite populations, when gene birth occurred, it was faster in smaller populations (Table.2(row2), Fig3, FigS4). We anticipate that the actual rate of gene birth in different organisms should scale according to features such as its generation time and mutation rate (Fig1(B), FigS1).

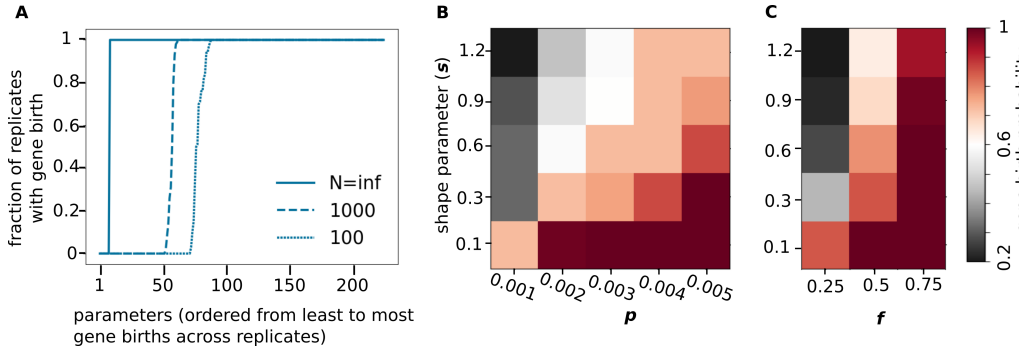


FIGURE 2 Probability of gene birth. (A) The fraction of replicate systems at different parameter values that achieve gene birth. See also FigS4. (B,C) Trade-off between the shape parameter, s (B) size of beneficial mutations, p and (C) the frequency of beneficial mutations, f. We show results here for populations of size N = 1,000. See FigS5 for N = 100, ∞. In the heatmaps, rows indicate values of the shape parameter and columns indicate values of parameters p and f, respectively. Colors indicate the fractions of systems with gene birth as shown in the colorbar. See FigS6 for effect of parameters on gene birth probability in populations of different sizes.

TABLE 2 Dynamics of gene birth for a genic threshold of 0.1. (Row1) The fraction of parameters out of 225 in which gene birth was observed. For finite populations, this is the number of parameters where gene birth occurred in at least one out of 100 replicate populations. (Row2) Average time to gene birth.

	N=100	N=1000	N=∞
Parameters with observed gene birth	68.0%	76.9%	96.4%
Mean time-steps to gene birth	380.7	452.4	357.7

Rare, large beneficial mutations suffice for gene birth

Reasonably, gene birth in the model was more likely when the frequency f and average size p of beneficial mutations are higher, and the size of deleterious mutations n is lower (FigS6). In addition, we find that the shape parameter s plays a crucial role: gene birth occurred 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) (Fig2(B,C), see also FigS5). This suggests that large-effect beneficial mutations are sufficient for gene birth, even when they are rare.

Gene birth is practically inevitable under DFE fluctuations

The DFE of a genomic locus is unlikely to remain the same over periods of millions of years that we simulate here. Broadly, there are three ways in which the DFE shifts in nature:

- Over relatively short periods of time, and under fairly constant environments, organisms experience 'diminishing returns epistasis', whereby the fitness gains due to beneficial mutations are smaller in relatively fit individuals than unfit individuals. Therefore, diminishing returns epistasis is likely to lead to decreased fitness gains along adaptive trajectories [18]. In terms of the model, this would look like the DFE parameter p reducing over time as fitter mutants undergo fixation in populations.
- Over longer periods, environmental changes could lead to DFE variations due to changes in the magnitude of mutation effects, maybe even switching the sign of some mutations from beneficial to deleterious and vice-versa [19].
- The DFE can also change because of changes in frequencies of different types of mutations, such as transitions or transversions. This can happen when there is a

shift in mutational biases, such as in mutator strains in *E. coli* [10].

Here, we test how fluctuations in the DFE could effect gene birth. We model DFE fluctuations by allowing one of the parameters of the DFE to shift to a neighbouring value at each time-step. For each initial parameter set, we run 10 replicate populations.

Now, the time-scale of each step in our model being of the order of hundreds to thousands of years, we do not expect the short-term effects of diminishing returns epistasis to persistently affect gene birth. Among the long-term fluctuations, we anticipate that environmental changes affect not only the DFE of new mutations, but also fitness contributions of preexisting genes. Such considerations are beyond the scope of the current model. Thus, the fluctuations we test here best represent shifts in mutational biases, which are expected to only affect the DFE of new mutations.

Studies indicate that shifts in mutational bias generally increase the rate of beneficial mutations [10]. But we allow fluctuations that decrease beneficial mutations; these potentially represent depletion of easily accessible beneficial mutations when the same mutational bias operates over long periods.

We find that allowing DFE parameters to fluctuate makes gene birth almost inevitable in systems of all population sizes tested, and regardless of initial parameters. In Fig3 (A,B,C), the grey points represent parameter values that did not lead to gene birth in any replicate population with static DFE. But under fluctuating DFE, gene birth occurs in all populations. For a large proportion of parameters, gene birth was faster, and gene birth times fall in a narrow range under fluctuating DFE (Fig3, see also FigS7).

To understand this high probability of gene birth, we notice that although each time-step involves only a small change in DFE parameters, this scheme of fluctuations uniformly samples all available DFEs. Over the 2,500 time-steps of the simulation, irrespective of initial parameters, almost all parameters are visited at least once (FigS8,9). Given that most DFEs are already highly conducive to gene birth (Fig2(A)), fluctuations in DFE allow all populations to spend substantial amounts of time in highly permissive parameters.

This in turn implies that gene birth depends on the duration a population spends in parameters conducive to gene birth, and is robust to the history of succession of these parameters. In other words, these systems are effectively equivalent irrespective of initial parameter values, thus also explaining the narrow distribution of time to gene birth.

Altogether, our analysis of the evolutionarily plausible scenario where the DFE of loci is allowed to change over time bolsters the view that *de novo* gene birth is a widespread process.

Discussion

The study of *de novo* genes draws attention to problems with our very conception of genes [1]. It is increasingly apparent that organisms adapt not only by optimizing preexisting genetic modules to environmental parameters, but also through the invention of new genes, which exist transiently [20]. How often genes are born *de novo* , the time-scale of gene birth, and the time for which they persist, are therefore important questions that are fundamental to our understanding of adaptive evolution.

We use a simple mathematical model to specifically illustrate the process by which beneficial mutations accumulate at non-genic loci and turn them into genes. In our model, *de novo* gene birth is a prevalent process, and particularly, rare, large-effect mutations play a major role in facilitating gene birth. Depending on the organism, a simulation of the model lasts for on the order of 10^5-10^6 years, and gene birth is highly likely within this time frame for a large range of parameters. Thus, our results lead us to put forth the hypothesis that that *de novo* gene birth is an unavoidable consequence of the constant adaptive evolution of non-genic loci.

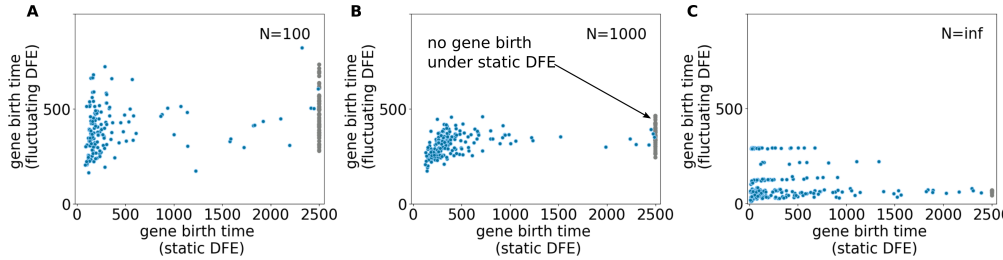


FIGURE 3 Scatter plots comparing gene birth time in systems with static vs. fluctuating DFE. For each point, the parameters for the static DFE and the initial parameters for fluctuating DFE are the same. The x-axis represents average time of gene birth across 100 replicate systems with static DFE, and the y-axis represents the average time of gene birth across 10 replicate systems when the DFE fluctuates, for (A) N = 100, (B) N = 1,000, and (C) N = ∞. The grey points that accumulate at x = 2,500 represent parameter values at which gene birth did not occur in static DFE systems. See also FigS7,8,9

Here, we use a coarse-grained definition of a gene, represented only by its fitness contribution. While this approach is immensely simple, it comes with trade-offs; firstly, birth of specific types of genes, such as anti-freeze proteins [21] or miRNA [22], might display very different dynamics. Particularly, emergence of function in protein coding *de novo* genes is very likely linked to the evolution of structural features [23]. Whereas the model is agnostic to the molecular mechanisms by which functionality is achieved.

On the other hand, this feature of the model is also useful, because in reality it is difficult to identify the emerging function of a *de novo* gene *a priori*, since we cannot envision what new factor in its environment the organism is next going to leverage. That is, new genes and their functions can only be identified with certainty post their establishment. But the evolutionary history of well-known *de novo* genes [21], and genomic studies that identify 'proto-genes' [24] indicate that the fitness contribution of new genes is built up gradually. In this scenario, this framework allows a glimpse into the dynamics of gene birth irrespective of the exact function of the new gene.

Secondly, our model cannot capture non-adaptive aspects of genome evolution. For example, we cannot test here the role of pre-adaptation: a non-adaptive process that leads to sequences evolving away from harmful phenotypes, but does not prescribe whether and how sequences gain new functions [25,26].

Nevertheless, the present framework allows us to base our assumptions on multitudes of experiments that measure organismal fitness in terms of an easily accessible, universal quantity: the relative growth rate. Numerous studies measure the effect of mutations on organismal fitness in terms of the DFE, which allows us to investigate the contribution of spontaneous mutations towards gene birth.

We use the DFE as a one-dimensional proxy for the various properties of genomic loci where *de novo* gene birth is observed, such as high GC content, presence of bidirectional promoters, and presence of meiotic hotspots [4]. For example, we imagine that at transcriptionally permissive genomic sequences, such as bidirectional promoters and nucleosome-free meiotic hotspots, any mutations are more likely to be expressed, and therefore are more likely to display DFE with higher values of p and n parameters. Moreover, the fitness contribution of any one locus necessarily depends on the cellular context through its interactions with other genes [27]. In this sense, the DFE also subsumes properties of the gene interaction network.

Ideally, a model of gene birth would describe the evolution of both expression level and functionality. In this

sense, currently available DFE measurements are limited in that they do not disentangle the contributions to a mutation's fitness effect due to changes in expression level versus changes in adaptive value. Such data would yield important insights into the dynamics of gene birth. For example, the birth of the *de novo* gene poldi in the house mouse is attributed mostly to mutations in regulatory regions that led to increased expression [28]. More generally, the availability of such data could help resolve whether the prevalent mode of birth of protein coding *de novo* genes is 'expression first', where loci are pervasively expressed and evolve functional features subsequently, or 'ORF first' [1], where loci that already possess ORFs gain expression [29].

The present results provoke a natural question: why do we not see many more *de novo* genes? While gene birth is likely to be a frequent and continuous process, the number of genes in a species is observed to remain fairly constant [29]. One reason for this apparent discrepancy is that recently emerged genes, typically with low expression levels, are much more likely to die than older genes, that are consistently expressed [22,20]. A second reason is the inability of current techniques to reliably detect *de novo* genes: new genes differ from established genes in sequence properties, such as length, disordered regions etc. [6]. Therefore, computational tools that are used to identify *de novo* genes, which are based on sequence properties learnt from established genes, under-count new genes. Potentially, these discrepancies could be resolved by replacing *de novo* gene identification methods that rely solely on sequence properties by methods that measure their fitness contributions. For example in [24], the fitness contributions of a chosen set of newly emerging genes are measured under conditions of gene disruption and over-expression. And in [22], the fitness effect of ablating *de novo* genes was measured to examine how new genes die. We envisage that large-scale measurements that employ such techniques and similar approaches could shed light on the process of emergence of *de novo* genes.

[1] Open Reading Frames (ORFs) are genomic sequences encoding a sequence of amino acids uninterrupted by stop codons; all cellular proteins are expressed from ORFs.

Methods

0.1 | Surveying the space of DFEs in populations of various sizes

We scan across DFEs with $p = [0.001, 0.002, 0.003, 0.004, 0.005]$, $f = [0.25, 0.5, 0.75]$, $n = [0.001, 0.005, 0.01]$ and $s = [0.1, 0.3, 0.6, 0.9, 1.2]$. And we look at populations of sizes $N = [100, 1000, \infty]$. For each parameter set, we simulate 100 replicate systems with finite population sizes, and simulate once for the infinite populations. In all, we look at 45,225 systems. We additionally look at fixation dynamics for 22,500 populations of size N = 5,000 (100 replicates across all parameter values). All codes used to generate and analyze data are written in Python3.6.

0.2 | Method to update population fitness for finite populations

For a population of size N, fitness of individuals at time-step t are stored the vector $F_t \in \mathbb{R}^{N\times 1}$, where the fitness of some individual i is $F_t(i)$. Now, only individuals with fitness > -1 are viable, and capable of producing progeny.

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

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

$$\text{and, } \mathrm{relfit}_t(i) = 0, \qquad \forall i \text{ s.t. } F_t(i) \leq -1$$

Let $\mathrm{Anc}_{t+1} \in \mathbb{N}^{NX1}$ 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 $t+1$. We define cumfit_t as the cumulative sum of relfit_t, and $\mathrm{Anc}_{t+1}(i) = \min(\{j \mid \mathrm{cumfit}_t(j) \geq U(0,1)\})$. Here $U(0,1)$ is a uniform random number between 0 and 1. This method ensures that the probability of choosing an individual is proportional to its fitness.

Progeny of the current population incur mutations. The values of fitness effects of mutations incurred by each individual at time-step t is stored in $\mathrm{mut}_t \in \mathbb{R}^{N\times 1}$, where

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

Here $\Gamma(\kappa,\theta)$ represents a number drawn from the gamma distribution with shape parameter κ and scale parameter θ, and $\mathrm{Ber}(p)$ is the Bernoulli random variable which equals 1 with probability p. The updated fitness levels of the population is then given by $F_{t+1}(i) = F_t(\mathrm{Anc}_{t+1}(i)) + \mathrm{mut}_t(i)$.

0.3 | Tracing ancestry and finding the fitness of the fixed mutation

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 $n = 1, 2, 3 \ldots t_0$, 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 $i \in X_{t-t0}$.

0.4 | Method to update population fitness for infinite populations

In order to look at the evolution of fitness in infinite populations, we fix a reasonable bound, $F_{\max} >> \text{genic-threshold}$, on the maximum value of fitness, and use $F_{\min} = -F_{\max}$ as the minimum value of fitness. We also discretize the fitness values into levels F^i separated by intervals of size δ, such that $F^{i+1} - F^i = \delta$.

Let $\mathrm{mut}(i \to j)$ be the probability density of a mutation of effect size $F^j - F^i$, where

$$\mathrm{mut}(i \to j) = f \cdot \Pr\left[\Gamma\left(\frac{p}{s}, s\right) = F^j - F^i\right], \qquad \text{if } F^j > F^i$$
$$\mathrm{mut}(i \to j) = (1-f) \cdot \Pr\left[\Gamma\left(\frac{n}{s}, s\right) = F^i - F^j\right], \qquad \text{if } F^i > F^j$$

The evolution of population fitness is then described by a transition matrix T, where

$$T(i,j) = \frac{\mathrm{mut}(i \to j)}{\sum_{k=-F_{\max}}^{F_{\max}} \mathrm{mut}(i \to k)}, \qquad \text{for } F^i > -1$$
$$\text{and,} \qquad T(i,j) = 0, \qquad \text{for } F^i \leq -1$$

Let the probability of occupancy of any fitness value F^j at time-step t be P_t^j. Here, relative fitness levels can be calculated as $\mathrm{relfit}^i = 1 - F^i - F_{\min}$. Then,

$$P_{t+1}^j = \frac{\sum_{i=-F_{\max}}^{F_{\max}} \left(P_t^i \cdot \mathrm{relfit}^i \cdot T(i,j)\right)}{\sum_{i=-F_{\max}}^{F_{\max}} \left(P_t^i \cdot \mathrm{relfit}^i\right)}\,.$$

references

1. Portin P, Wilkins A. The evolving definition of the term "gene". *Genetics* 2017; 205(4): 1353–1364.

2. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nature Reviews Genetics* 2011; 12(10): 692–702.

3. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* 2020; 9: e53500.

4. Vakirlis N, Hebert AS, Opulente DA, et al. A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution* 2018; 35(3): 631–645.

5. Jensen TH, Jacquier A, Libri D. Dealing with pervasive transcription. *Molecular cell* 2013; 52(4): 473–484.

6. Van Oss SB, Carvunis AR. De novo gene birth. *PLoS genetics* 2019; 15(5): e1008160.

7. Carvunis AR, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature* 2012; 487(7407): 370–374.

8. Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T. Can an arbitrary sequence evolve towards acquiring a biological function?. *Journal of molecular evolution* 2003; 56(2): 162–168.

9. Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. *Nature communications* 2018; 9(1): 1–10.

10. Sane M, Diwan GD, Bhat BA, Wahl LM, Agashe D. Shifts in mutation spectra enhance access to beneficial mutations. *bioRxiv* 2020.

11. Böndel KB, Kraemer SA, Samuels T, et al. Inferring the distribution of fitness effects of spontaneous mutations in Chlamydomonas reinhardtii. *PLoS biology* 2019; 17(6): e3000192.

12. Racimo F, Schraiber JG. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genet* 2014; 10(11): e1004697.

13. Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences* 2017; 114(17): 4465–4470.

14. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* 2007; 8(8): 610–618.

15. Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. Extensive de novo mutation rate variation between individuals and across the genome of Chlamydomonas reinhardtii. *Genome Research* 2015; 25(11): 1739–1749.

16. Gillespie JH. *Population genetics: a concise guide*. JHU press . 2004.

17. Díaz J, Goldberg LA, Mertzios GB, Richerby D, Serna M, Spirakis PG. Approximating fixation probabilities in the generalized moran process. *Algorithmica* 2014; 69(1): 78–91.

18. Wünsche A, Dinh DM, Satterwhite RS, Arenas CD, Stoebel DM, Cooper TF. Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. *Nature Ecology & Evolution* 2017; 1(4): 1–6.

19. Das SG, Direito SO, Waclaw B, Allen RJ, Krug J. Predictable properties of fitness landscapes induced by adaptational tradeoffs. *Elife* 2020; 9: e55155.

20. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. *elife* 2014; 3: e01311.

21. Zhuang X, Yang C, Murphy KR, Cheng CHC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proceedings of the National Academy of Sciences* 2019; 116(10): 4400–4405.

22. Lu GA, Zhao Y, Liufu Z, Wu CI. On the possibility of death of new genes–evidence from the deletion of de novo microRNAs. *BMC genomics* 2018; 19(1): 1–8.

23. Papadopoulos C, Callebaut I, Gelly JC, et al. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *bioRxiv* 2021.

24. Vakirlis N, Acar O, Hsu B, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications* 2020; 11(1): 1–18.

25. Masel J. Cryptic genetic variation is enriched for potential adaptations. *Genetics* 2006; 172(3): 1985–1991.

26. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature ecology & evolution* 2017; 1(6): 1–6.

27. Wei X, Zhang J. Patterns and mechanisms of diminishing returns from beneficial mutations. *Molecular biology and evolution* 2019; 36(5): 1008–1021.

28. Heinen TJ, Staubach F, Häming D, Tautz D. Emergence of a new gene from an intergenic region. *Current biology* 2009; 19(18): 1527–1531.

29. Schlötterer C. Genes from scratch–the evolutionary fate of de novo genes. *Trends in Genetics* 2015; 31(4): 215–219.

30. Farlow A, Long H, Arnoux S, et al. The spontaneous mutation rate in the fission yeast Schizosaccharomyces pombe. *Genetics* 2015; 201(2): 737–744.

31. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome research* 2009; 19(7): 1195–1201.

32. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences* 2012; 109(41): E2774–E2783.

33. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences* 2008; 105(46): 17878–17883.

34. Long H, Winter DJ, Chang AYC, et al. Low base-substitution mutation rate in the germline genome of the ciliate Tetrahymena thermophila. *Genome biology and evolution* 2016; 8(12): 3629–3639.

35. Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. *Genetics* 2012; 192(4): 1447–1454.

36. Ossowski S, Schneeberger K, Lucas-Lledó JI, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *science* 2010; 327(5961): 92–94.

37. Saxer G, Havlak P, Fox SA, et al. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba Dictyostelium discoideum. *Plos One* 2012.

38. Sung W, Tucker AE, Doak TG, Choi E, Thomas WK, Lynch M. Extraordinary genome stability in the ciliate Paramecium tetraurelia. *Proceedings of the National Academy of Sciences* 2012; 109(47): 19339–19344.

39. Zhu YO, Siegal ML, Hall DW, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences* 2014; 111(22): E2310–E2318.

40. Cassidy-Hanley DM. Tetrahymena in the laboratory: strain resources, methods for culture, maintenance, and storage. *Methods in cell biology* 2012; 109: 237–276.

41. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* 2010; 38(suppl_1): D750–D753.

42. Petersen J, Russell P. Growth and the environment of Schizosaccharomyces pombe. *Cold Spring Harbor Protocols* 2016; 2016(3): pdb–top079764.

43. Harris EH. *The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and Its Laboratory Use: Volume 1*. 1. Academic press . 2009.

44. Ishida M, Hori M. Improved isolation method to establish axenic strains of Paramecium. *Japanese Journal of Protozoology* 2017; 50(1-2): 1–14.

45. Fey P, Kowal AS, Gaudet P, Pilcher KE, Chisholm RL. Protocols for growth and development of Dictyostelium discoideum. *Nature protocols* 2007; 2(6): 1307–1316.

46. Silva RR, Moraes CA, Bessan J, Vanetti MCD. Validation of a predictive model describing growth of Salmonella in enteral feeds. *Brazilian Journal of Microbiology* 2009; 40(1): 149–154.

47. Fernández-Moreno MA, Farr CL, Kaguni LS, Garesse R. Drosophila melanogaster as a model system to study mitochondrial biology. In: Springer. 2007 (pp. 33–49).

48. Koornneef M, Scheres B. Arabidopsis thaliana as an experimental organism. *e LS* 2001.

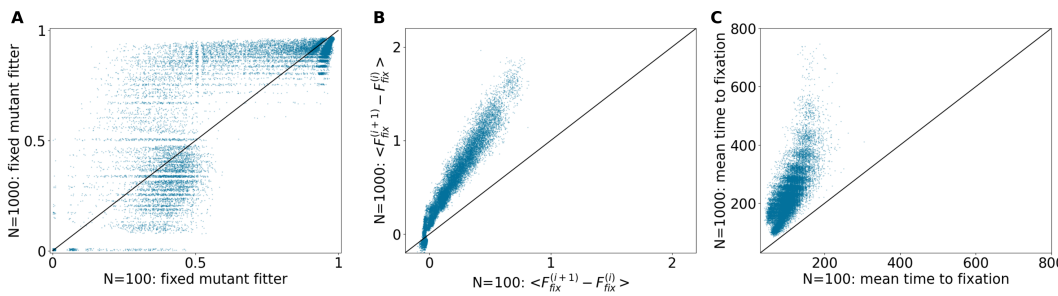
Acknowledgements

This work was funded by the Institute for Basic Science, Grant IBS-R020.

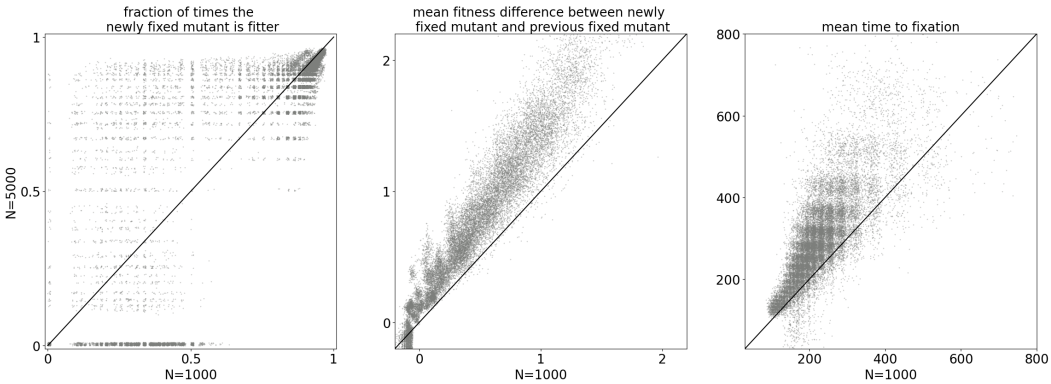
Supplementary information

Species	Generation time	Mutation rate (per base pair per generation)	Model time-step
T.thermophila	2 hrs	$7.6*10^{-12}$	$3 * 10^5$ yrs
S.cerevisiae	90 min	$1.6*10^{-10}$	$1.1 * 10^4$ yrs
E.coli	20 min	$2*10^{-10}$	$1.9 * 10^3$ yrs
S.pombe	2 hrs	$2*10^{-10}$	$1.1 * 10^4$ yrs
C.reinhardtii	6 hrs	$3*10^{-10}$	$2.3 * 10^4$ yrs
P.tetraurelia	8 hrs	$2*10^{-11}$	$4.6 * 10^5$ yrs
D.discoideum	4 hrs	$2.9*10^{-11}$	$1.6 * 10^5$ yrs
S.typhimurium	21 min	$3*10^{-9}$	133 yrs
D.melanogaster	10 days	$7*10^{-9}$	$7.8 * 10^4$ yrs
A.thaliana	6 weeks	$9*10^{-9}$	$2.4 * 10^4$ yrs

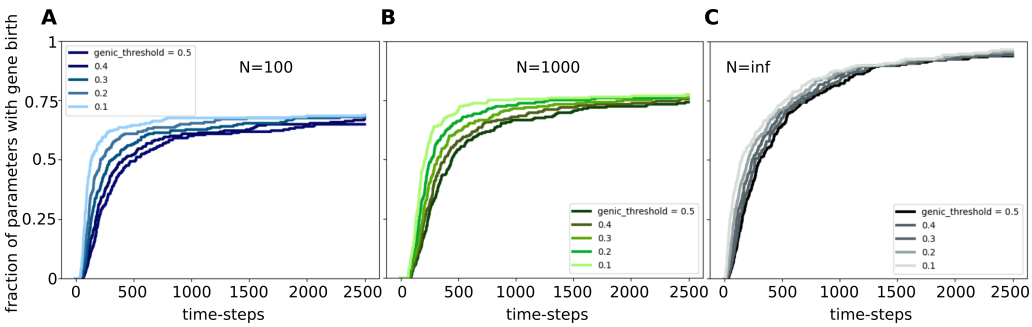
FIGS 1 Estimates of one model time-step in various species. The following references were used to obtain mutation rates:[30,31,32,33,34,35,36,37,38,39]. And the following references were used for generation times:[40,41,42,43,44,45,46,47,48]. Most references give a range of values or multiple values of mutation rates and generation times depending on the culture conditions used. In such cases, we picked the average reported value at a single culture condition that is consistent between the studies reporting mutation rate and generation time in any one organism. The model time-step calculated in column 3 is the time it takes for 1 mutation to occur in a locus of 100 bp.



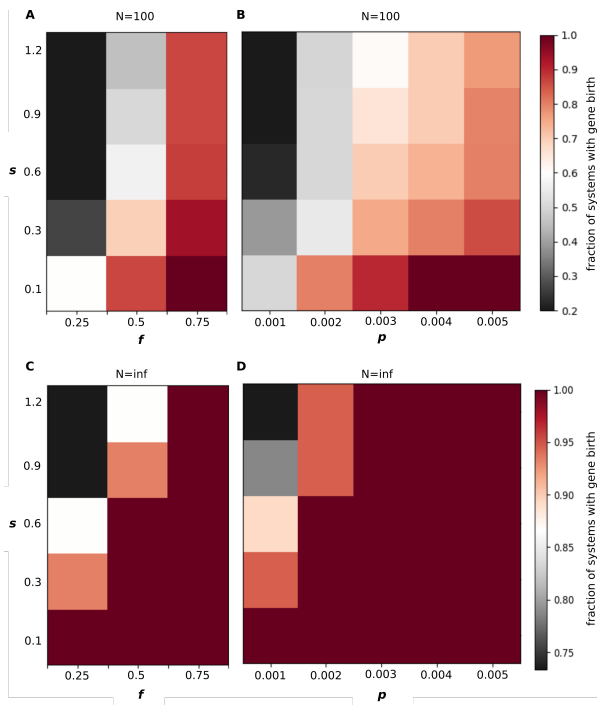
FIGS 2 **Scatter plots comparing fixation dynamics in** $N = 100$ **and** 1,000 **populations.** For each population size, 22,500 systems were analysed across all parameter values. For each point, the x-coordinate represents an N = 100 and the y-coordinate represents an N= 1,000 population that have the same parameter values. (A) Fraction of times fixation of fitter mutant occurred, (B) mean fitness difference between current and previous fixed mutant, $\langle \Delta F_{\text{fix}} \rangle = \langle F_{\text{fix}}^{i+1} - F_{\text{fix}}^{i} \rangle$, where F_{fix}^{i} is the fitness of the ith mutant that fixed in the population, and (C) mean fixation time across 2,500 time steps.



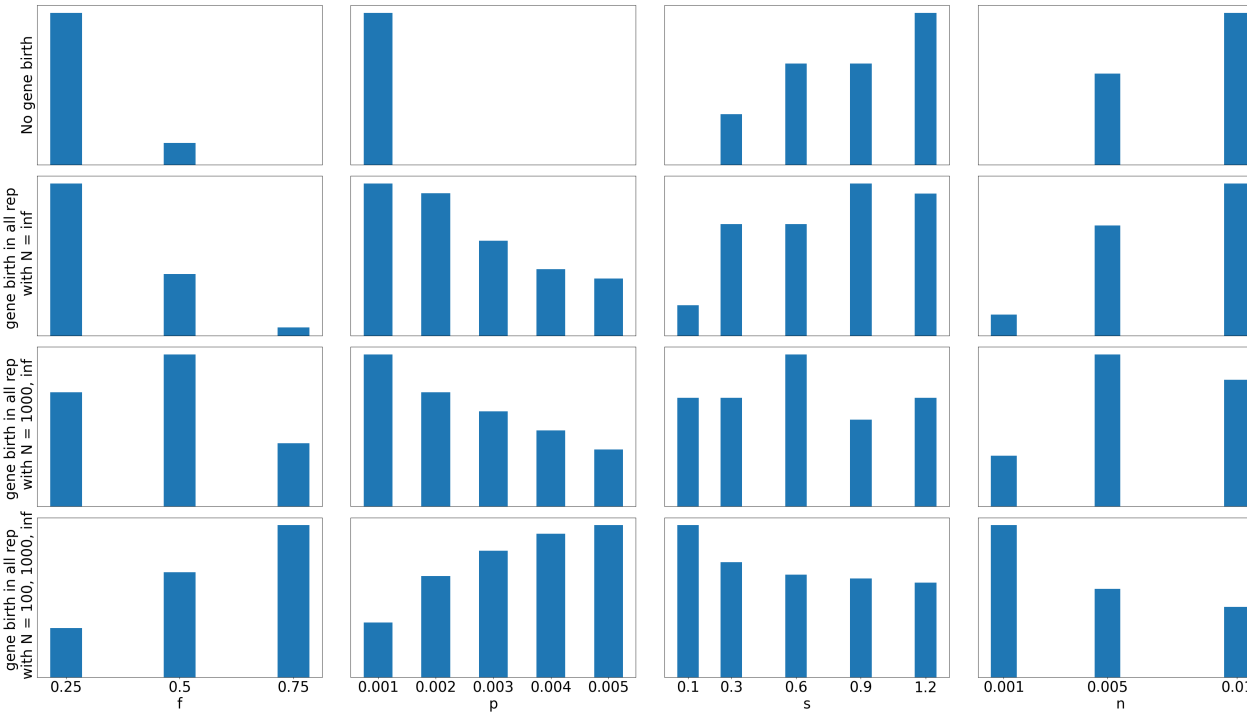
FIGS 3 Scatter plots comparing fixation dynamics in N = 1000 and 5000 populations. For each population size, 22,500 systems were analysed across all parameter values. The x and y value for each point represents (A) Fraction of times fixation of fitter mutant occurred, (B) mean fitness difference between current and previous fixed mutant, (C) Mean fixation time across 2500 time steps, for an N = 1000 and an N = 5000 systems that have equal parameter values



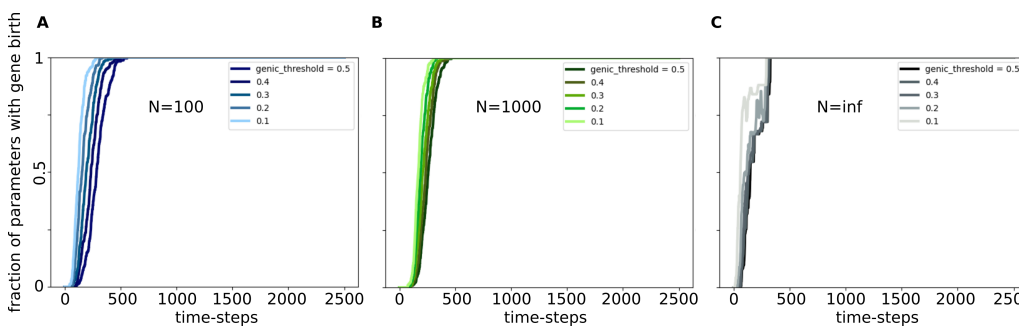
FIGS 4 In (A,B,C) the x-axis represents time-steps and the y-axis represents the fraction of parameters tested for which gene birth occurred at least in 1 of the 100 replicate populations tested. Population sizes: (A) N=100, (B) N=1000, (C) infinite population.



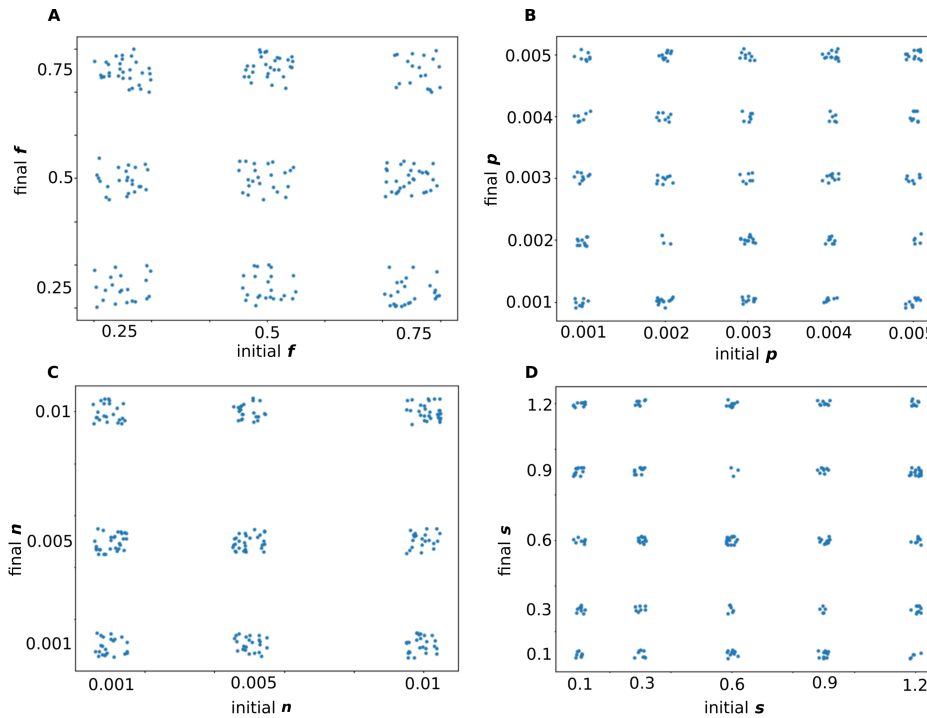
FIGS 5 Trade-off between the shape parameter and frequency and size of beneficial mutations. We show results here for populations of sizes (A,B) N = 100 and (C,D) N = inf. In the heatmaps of rows indicate values of the shape parameter, columns indicate values of (A,C) fraction of beneficial mutation *f*, (B,D) mean size of beneficial mutations *p*. Colors indicate the fractions of systems with gene birth as indicated by the colorbar.



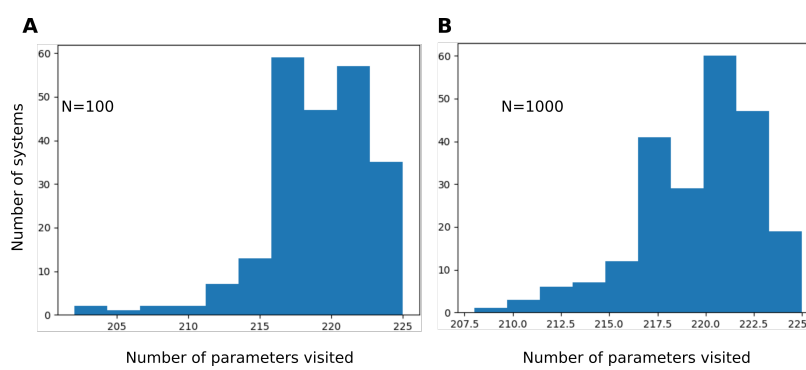
FIGS 6 Effect of model parameters on gene birth probability. Each row contains histograms for parameter values at which *first row*: no gene birth occurred in any population, *second row*: gene birth occurred in all 100 replicates only in N = inf populations, *third row*: gene birth occurred in all 100 replicates only in N = 1000 and N = inf populations, *fourth row*: gene birth occurred in all 100 replicates across all N = 100, 1000, inf. Columns correspond to different parameters, first column: *f*, second column: *p*, third column: *s*, fourth column: *n*



FIGS 7 All parameters lead to gene birth within 2500 time-steps under the fluctuating DFE regime. In (A,B,C) the x-axis represents time-steps and the y-axis represents the fraction of parameters tested for which gene birth occurred at least in 1 of the 100 replicate populations tested. Population sizes: (A) N=100, (B) N=1000, (C) infinite population.



FIGS 8 Scatter plots for initial and final values of parameters under fluctuating DFE regime. Each point represents a distinct N = 1000 population. 2250 systems were analysed for this figure. Noise has been added to make the density of points more apparent. (A) fraction of beneficial mutations *f*, (B) mean size of beneficial mutations *p*, (C) mean size of deleterious mutations *n*, (D) shape parameter *s*



FIGS 9 Histograms for number of distinct parameters visited by populations under the fluctuating DFE regime. 2250 systems were analysed for each histogram. (A) N = 100, (B) N = 1000