Ease of *de novo* gene birth through spontaneous mutations predicted in a parsimonious model

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

***For correspondence:**
somyamn@gmail.com (SM);
tsvitlusty@gmail.com (TT)

[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, Republic of Korea; [2]Departments of Physics and Chemistry, UNIST, Ulsan 44919, Republic of Korea

Abstract Contrary to long-held views, recent evidence indicates that *de novo* birth of genes is not only possible, but is surprisingly prevalent: a substantial fraction of eukaryotic genomes are composed of orphan genes, which show no homology with any conserved genes. And a remarkably large proportion of orphan genes likely originated *de novo* from non-genic regions. Here, using a parsimonious mathematical model, we investigate the probability and timescale of *de novo* gene birth due to spontaneous mutations. We trace how an initially non-genic locus accumulates beneficial mutations to become a gene. We sample across a wide range of biologically feasible distributions of fitness effects (DFE) of mutations, and calculate the conditions conducive to gene birth. We find that in a time frame of millions of years, gene birth is highly likely for a wide range of DFEs. Moreover, when we allow DFEs to fluctuate, which is expected given the long time frame, gene birth in the model becomes practically inevitable. This supports the idea that gene birth is a ubiquitous process, and should occur in a wide variety of organisms. Our results also demonstrate that intergenic regions are not inactive and silent but are more like dynamic storehouses of potential genes.

Introduction

Most generally, a gene can be described as a genomic sequence specifying products, which through regulatory interactions with other genes and the environment, yields phenotypic traits (***Portin and Wilkins, 2017***). It is a long-held belief that a basic set of 'founder genes', numbering some thousands, evolved a long time ago, and all new genes are exclusively derived from these founder genes. In contrast to this 'genes come from genes' picture, recent evidence indicates that 10 – 30 % of all genes across eukaryotic genomes are orphan genes, for which no homology can be detected with any established, conserved genes (***Tautz and Domazet-Lošo, 2011***). 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 (***Vakirlis et al., 2020b***). Moreover, evidence from phylostratigraphy suggests that such new genes are generated continuously (***Tautz and Domazet-Lošo, 2011***).

Patterns observed in studies on well-known *de novo* genes and putative *de novo* genes shed light on possible mechanisms of gene birth: new genes are born preferably in genomic regions with high GC content and near meiotic hotspots. In animals, new genes are more likely to be expressed in the brain and testis (***Vakirlis et al., 2018***). Interestingly, these cells and genomic regions are also especially prone to pervasive transcription (***Jensen et al., 2013***). These patterns suggest an intriguing hypothesis (***Van Oss and Carvunis, 2019***), whereby non-genic loci, made visible to nat-

ural selection by pervasive expression, can be driven to evolve in two ways: the sequences could evolve towards reducing deleterious traits, such as propensity to aggregate, in a process called pre-adaptation (***Masel, 2006***; ***Wilson et al., 2017***); alternatively, or additionally, the sequences can gain new functions that increase the fitness of the organism (***Carvunis et al., 2012***). Pre-adaptation is considered non-adaptive evolution since it ensures that expressed sequences remain harmless, but in itself does not prescribe how the expression of these sequences increase organismal fitness. Whereas gain of a new function is considered adaptive and leads to gene birth.

Two simple observations, taken together, lend support to such a mechanism of gene birth: first, random sequences can gain functionality, provided they are consistently expressed (***Hayashi et al., 2003***). And second, it was demonstrated that new promoters could easily evolve in *E. coli* (***Yona et al., 2018***). These studies highlight the possibility that non-genic sequences can, in stages, gain the hallmarks of genes: regulated expression, and subsequently, functionality.

In this work, we use a mathematical model to assess the expected frequency and time-scale of *de novo* gene birth. To this end, we look at gene birth in an essentially adaptive framework (***Orr, 2005***), with a key difference: instead of a highly fit wild-type ancestor, *de novo* gene birth begins, by definition, at loci with very low fitness. Here, instead of the emergence of some specified function, we consider the process by which a locus that initially provides no fitness advantage starts to accumulate beneficial mutations. We conjecture that this process reflects the early stages of *de novo* gene birth. Note that we do not constrain genes to only those that produce proteins; we consider both proteins and RNA to be products of gene expression.

Experimentally, fitness effects of spontaneous mutations are assessed through mutation accumulation studies. These studies directly (***Sane et al., 2020***), or indirectly (***Böndel et al., 2019***) allow inference of the fitness effects of single mutations, thereby yielding a distribution of fitness effects (DFE). Here, we sample such biologically realistic DFEs to assess the conditions under which spontaneous mutations, mainly single nucleotide mutations and short indels, can lead to gene birth. The different DFEs reflect different species (***Huber et al., 2017***), or different regions within the genome (***Racimo and Schraiber, 2014***).

Our description of a functional gene is minimal and does not concern what the function is or the mechanism through which the gene achieves its function. Our model of gene birth is general, and admittedly, birth of specific classes of genes, such as anti-freeze proteins (***Zhuang et al., 2019***) or miRNA (***Lu et al., 2018***), might display very different dynamics. At the same time, the model's simplicity allows us to leverage a wide array of data and reach biologically reasonable conclusions.

Our results indicate that gene birth is 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 compensate for beneficial mutations being small and infrequent on average. As expected, gene birth in the model was more frequent in larger populations. We also tested the more realistic scenario where the DFE of a genomic locus fluctuates over the long periods that we consider in the model. Surprisingly, we find that under these conditions, gene birth is practically inevitable and occurs at an even faster rate.

Gene birth model

We use the Wright-Fischer 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., ***Böndel et al. (2019)***). We describe a locus as *genic* if it consistently contributes a fitness advantage above a predetermined *genic threshold*.

Initially, the locus is non-genic, and the distribution of F_i is centered around 0. In each subsequent time-step $t+1$, the population is composed entirely of the offspring of individuals in the current time-step t (Fig.1(A)). The probability that an individual leaves an offspring is proportional

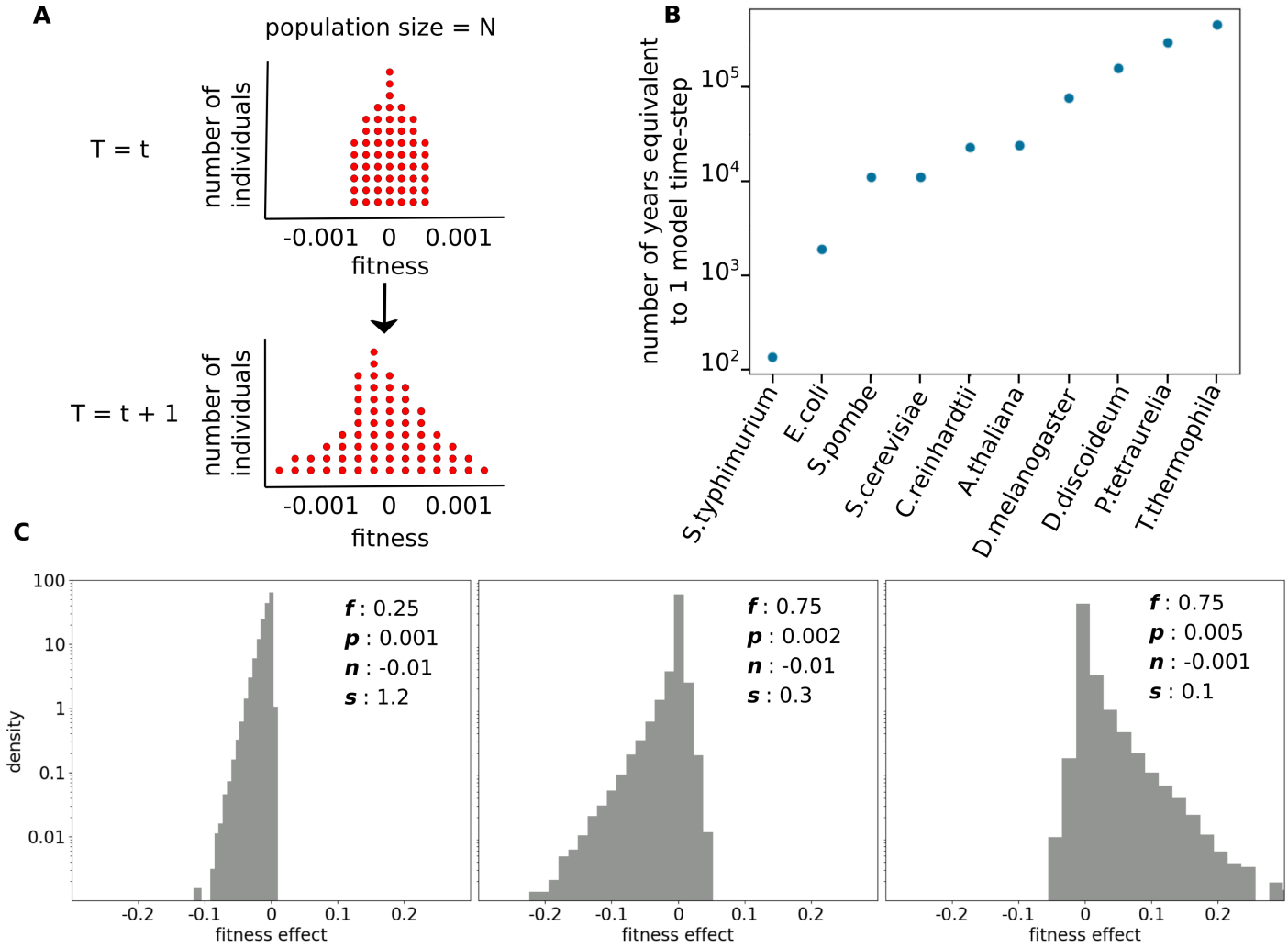

Figure 1. Time-scale and fitness effects of mutations in the model. (A) Model schematic. Each red dot represents the fitness contribution of the locus of interest in an individual of the population. At some time step $T = t + 1$, the population is replaced by offspring of the previous population at $T = t$. The fitness of an offspring depends on the fitness of its parent and the effect of any mutation incurred. (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 Fig1–Figure Supplement1 for calculations. (C) DFE for different model parameters. The left panel represents the DFE with the most deleterious and least beneficial mutations. The right panel represents the DFE with the most beneficial and least deleterious mutations sampled in this work. The center panel represents the DFE with parameters closest to those reported in (***Böndel et al., 2019***). Values of parameters used to construct these DFEs are given alongside each histogram.

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 (Fig.1(B), see also Fig1–Figure Supplement1).

The fitness effects of mutations are drawn from the characteristic DFE for the locus (Fig.1(C)). Following ***Böndel et al.*** (***2019***), we 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 shape parameters are more long-tailed. The mutation types we deal with in the model are those that were included in ***Böndel et al.*** (***2019***), which were single-nucleotide mutations and short indels (on average ≤ 10 bp) (***Ness et al., 2015***).

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). We 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 (see Method to update population fitness 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.

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.

Results

Small populations fix fewer beneficial mutations faster

In our model, we look at fixation dynamics in populations of size N = 100, 1,000 and 5,000. It is well-known in population dynamics that small populations are generally subject to stronger genetic drift, which makes it harder for fitter mutants to fix (***Gillespie*** (***2004***), chapter 2). Additionally, mathematical models of the Moran process indicate that fitter mutants have a lower fixation probability and a shorter fixation time in smaller populations (***Díaz et al., 2014***). Consistently, we find that the probability of fixation of fitter *de novo* mutants that arise in smaller populations is lower (Fig.2(A), Table.1(row1), Fig2–Figure Supplement1(A)). Moreover, 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 (Fig.2(B), Fig2–Figure Supplement1(B)). And fixation occurs much faster in $N = 100$ populations (Fig.2(C), Table.1(row2), Fig2–Figure Supplement1(C)).

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

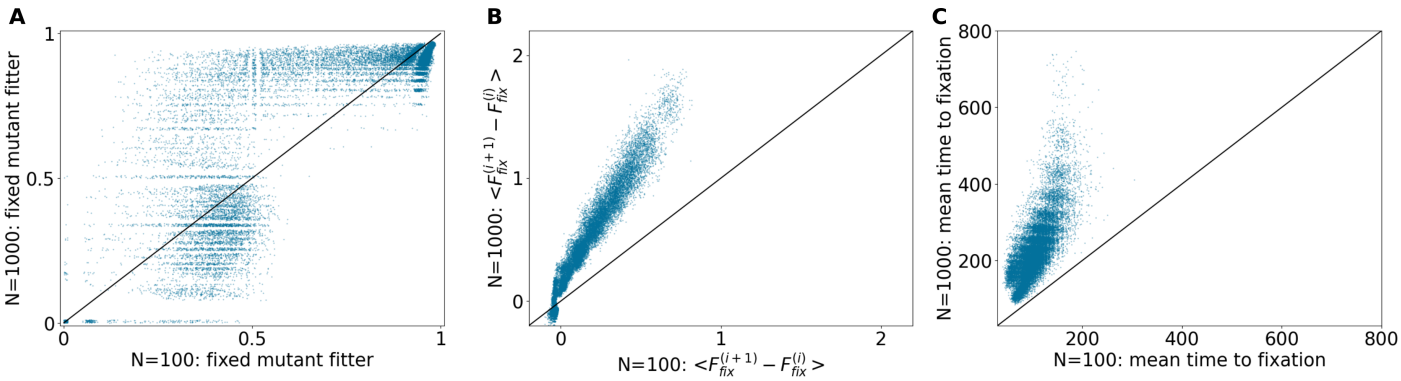

Figure 2. Scatter plots comparing fixation dynamics in $N = 100$ and 1,000 populations. For each population size, 22,500 systems were analysed across all parameter values. For each point, the x-coordinate represents an N = 100 and the y-coordinate represents an N= 1,000 population that have the same 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. See Fig2–Figure Supplement1 for scatter plots comparing N = 1,000 and 5,000.

Most parameters allow gene birth

Surprisingly, a majority of parameters in our survey are conducive to gene birth (Fig.3(A)). This finding suggests that gene birth due to spontaneous mutations should be a universal process, and we should expect it to occur in a wide variety of organisms and across many distinct genomic regions. The rate of gene birth should scale according to features of the organisms such as its generation time and mutation rate (Fig.1(B), Fig1–Figure Supplement1). As expected, gene birth is more prevalent in larger populations (Table.2(row1)). All parameters that allow gene birth in a small population also allow it in larger populations. Although the number of permissive parameters decreases as the genic threshold is increased, the results remain qualitatively the same (Fig3–Figure Supplement1). Also, as expected from fixation time distributions for finite populations, when gene birth occurred, it was faster in smaller populations (Table.2(row2), Fig.4, Fig3–Figure Supplement1).

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-effect beneficial mutations are sufficient for gene birth

Across all population sizes, gene birth is more likely when the frequency f and average size p of beneficial mutations are higher, the size of deleterious mutations n is lower, and the DFE is more long-tailed (Fig3–Figure Supplement3). Although gene birth depends strongly on beneficial mutations, we find that it can still occur in populations with small values of parameters f and p, provided the DFE of mutations is long-tailed (i.e., small values of s) (Fig.3(B,C), see also Fig3–Figure Supplement2). 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 likely to change over time. In addition to sampling across different parameters, we also test how fluctuations in the DFE could effect gene birth. Broadly, there are three ways in which the DFE shifts in nature:

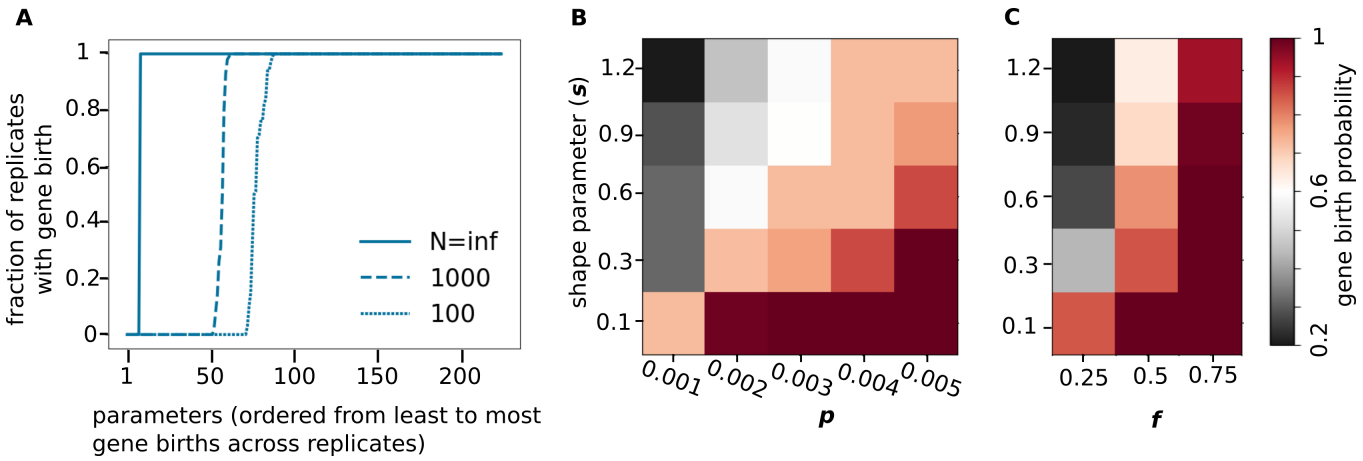

Figure 3. Probability of gene birth. (A) The fraction of replicate systems at different parameter values that achieve gene birth. See also Fig3–Figure Supplement1. (B,C) Trade-off between the shape parameter, s (B) size of beneficial mutations, p and (C) the frequency of beneficial mutations, f. We show results here for populations of size N = 1,000. See Fig3–Figure Supplement2 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 Fig3–Figure Supplement3 for effect of parameters on gene birth probability in populations of different sizes.

- Over relatively short periods of time, 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 (***Wünsche et al., 2017***). 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 (***Das et al., 2020***).
- 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* (***Sane et al., 2020***).

Since the time-scale of each step in our model is of the order of hundreds to thousands of years, we do not expect diminishing returns epistasis to persistently affect gene birth. And environmental changes are expected to change not only the DFE of new mutations, but also fitness contributions of preexisting genes, which is beyond the scope of the current model. Thus, the fluctuations we test in the model best represent shifts in mutational biases, which are expected to only affect the DFE of new mutations. This is modelled 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. While shifts in mutational bias are predicted to increase beneficial mutations (***Sane et al., 2020***), we also allow fluctuations that decrease beneficial mutations; these could represent a 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 Fig.4 (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 (Fig.4, see also Fig4–Figure Supplement1). To understand this high probability of gene birth under this scheme of DFE fluctuations, we notice that although each time-step involves only a small

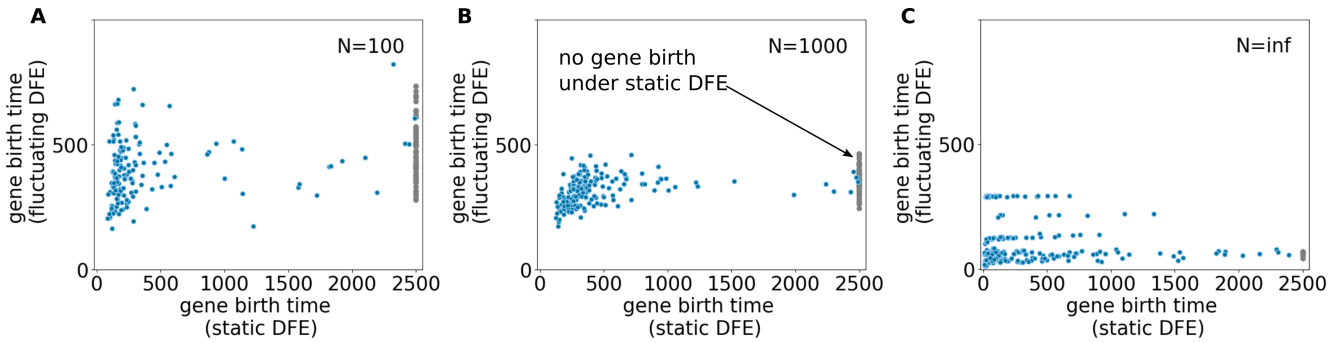

Figure 4. Scatter plots comparing gene birth time in systems with static vs. fluctuating DFE. For each point, the parameters for the static DFE and the initial parameters for fluctuating DFE are the same. The x-axis represents average time of gene birth across 100 replicate systems with static DFE, and 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 Fig4–Figure Supplement1,2,3

change in DFE parameters, almost all parameter sets are visited at least once over 2,500 time-steps, irrespective of initial parameters (Fig4–Figure Supplement2,3). Now, across N = 100, 1,000 populations, 60.9 % and 72.4 %, respectively, of all parameters led to gene birth in all 100 replicate systems (Fig.3(A)). And, on average, these are the fractions of time steps along the 2,500 time-step trajectory in which the fluctuating systems visited these high gene birth parameters. This indicates that this scheme of DFE fluctuations uniformly samples all available DFEs over long enough periods of time. The substantial time spent in highly permissive parameters explains the high incidence of gene birth. It also implies that these systems are effectively equivalent irrespective of initial parameter values, thus explaining the narrow distribution of time to gene birth. That is, gene birth depends on the duration a population spends in conditions conducive to gene birth, and is robust to the history of succession of these conditions. Altogether, this 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 what a gene is (***Portin and Wilkins, 2017***). It appears that organisms adapt not only by optimizing preexisting genetic modules to environmental parameters, but also through the invention of new genes, which exist transiently (***Palmieri et al., 2014***). 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 present here a simple mathematical model to study aspects of *de novo* gene birth. Principally, we investigate the probability and the expected time frame of gene birth under various conditions. Our conception of a 'gene' is coarse-grained, representing the gene only by its adaptive value. For instance, the emergence of function in *de novo* genes that express proteins is very likely linked to the evolution of structural features (***Papadopoulos et al., 2021***), but our model is agnostic to the molecular mechanisms by which functionality is achieved. Our approach is therefore immensely simple, but it comes with trade-offs; importantly, this framework cannot capture non-adaptive aspects of genome evolution. For example, we cannot test here the role of pre-adaptation, which is a non-adaptive process that leads to sequences evolving away from highly deleterious phenotypes, but does not prescribe whether and how sequences gain new functions (***Masel, 2006***). Nevertheless, an advantage of the present framework is the multitude of experimental studies 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.

Here we use the DFE as a one-dimensional proxy for the huge multidimensional space that describes the properties of genomic sequences. That is, we assume that the various properties of loci where *de novo* gene birth is observed, such as high GC content, presence of bidirectional promoters, and presence of meiotic hotspots (***Vakirlis et al., 2018***), all enter the system via their influence on the DFE of that genomic locus. 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. The DFE also subsumes properties of the gene interaction network since the fitness contribution of any one locus depends on the cellular context in which it is expressed (***Wei and Zhang, 2019***).

A limitation of DFE measurements, especially for the question of gene birth, is that current data do not disentangle the contributions to a mutation's fitness effect due to changes in expression level versus changes in adaptive value of the locus. 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, which led to increased expression (***Heinen et al., 2009***). More generally, the availability of such data could help resolve whether the prevalent mode of *de novo* birth of protein-producing genes is 'expression first' (loci are pervasively expressed and evolve functional features subsequently), or 'ORF first' (loci that already possess ORFs gain expression) (***Schlötterer, 2015***).

Our results give us a glimpse into the dynamics of gene birth irrespective of the exact function the new gene performs. This is useful, because it is difficult to identify *a priori* what new functions could arise in an organism, as 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 (***Zhuang et al., 2019***), and genomic studies that identify 'proto-genes' (***Vakirlis et al., 2020a***) indicate that the adaptive value of new genes is built up gradually. In this scenario, our model provides a means for independent theoretical validation of observations from genomics studies about the prevalence of *de novo* gene birth, and produces estimates for the time-scale of the emergence of new genes.

Our model supports the idea that *de novo* gene birth is a widespread and frequent process, and depending on the organism, the process of gene birth should take on the order of $10^5 - 10^6$ years. Particularly, our results suggest that rare, large-effect mutations that occur in the more long-tailed DFEs play a major role in facilitating gene birth. Also, under the regime of fluctuating DFE, which represents shifts in mutational biases, the likelihood of *de novo* gene birth increases even further. While large events, such as transposition, are very likely to occur in this time frame, and are known to be associated with *de novo* gene birth (***Ruiz-Orera et al., 2015***), our results suggest that small-scale spontaneous mutations, such as substitutions and small indels on their own are capable of pushing non-genic loci towards 'genehood'.

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 over long periods of time (***Schlötterer, 2015***). One reason for this discrepancy is that recently emerged genes with low expression levels are much more likely to die than older genes that are consistently expressed (***Lu et al., 2018***; ***Palmieri et al., 2014***). 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. (***Van Oss and Carvunis, 2019***). Therefore, computational tools that are used to identify *de novo* genes 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 (***Vakirlis et al., 2020a***), the fitness contributions of a chosen set of newly emerging genes are measured under conditions of gene disruption and over-expression. And in (***Lu et al., 2018***), the fitness effect of ablating *de novo* genes was measured to examine how new genes die. Using such techniques

and similar approaches, large-scale measurements of fitness effects of expression products can be employed to gain more insight into how *de novo* genes emerge.

Methods

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.

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 $\text{allfit}_t = \sum_j \left(1 + F_t(j) - \text{minfit}_t\right)$, 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

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

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 $t+1$. We define cumfit_t as the cumulative sum of relfit_t, and $\text{Anc}_{t+1}(i) = \min\left(\left\{j | \text{ cumfit}_t(j) \geq U(0,1)\right\}\right)$. 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 $\text{mut}_t \in \mathbb{R}^{N\times 1}$, where

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

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

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 $X_t = \{i, \forall i \in \text{Anc}_t\}$ be the set of unique ancestor identities. We then recursively find $X_{t-n} = \{i, \forall i \in \{\text{Anc}_{t-n}(j), \forall j \in X_{t-n+1}\}\}$ 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 $i \in X_{t-t0}$.

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

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

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

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

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 $\text{relfit}^i = 1 - F^i - F_{\min}$. Then,

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

Acknowledgments

This work was funded by the Institute for Basic Science, South Korea, Project Code IBS-R020-D1.

Competing interests

The authors declare that they have no competing interests.

References

Böndel KB, Kraemer SA, Samuels T, McClean D, Lachapelle J, Ness RW, Colegrave N, Keightley PD. Inferring the distribution of fitness effects of spontaneous mutations in Chlamydomonas reinhardtii. PLoS biology. 2019; 17(6):e3000192.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. Proto-genes and de novo gene birth. Nature. 2012; 487(7407):370–374.

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

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

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.

Farlow A, Long H, Arnoux S, Sung W, Doak TG, Nordborg M, Lynch M. The spontaneous mutation rate in the fission yeast Schizosaccharomyces pombe. Genetics. 2015; 201(2):737–744.

Fernández-Moreno MA, Farr CL, Kaguni LS, Garesse R. Drosophila melanogaster as a model system to study mitochondrial biology. In: *Mitochondria* Springer; 2007.p. 33–49.

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.

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

Harris EH. The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and Its Laboratory Use: Volume 1, vol. 1. Academic press; 2009.

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.

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

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.

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

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

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.

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

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.

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

Long H, Winter DJ, Chang AYC, Sung W, Wu SH, Balboa M, Azevedo RB, Cartwright RA, Lynch M, Zufall RA. Low base-substitution mutation rate in the germline genome of the ciliate Tetrahymena thermophila. Genome biology and evolution. 2016; 8(12):3629–3639.

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.

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

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.

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

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.

Orr HA. The genetic theory of adaptation: a brief history. Nature Reviews Genetics. 2005; 6(2):119–127.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. science. 2010; 327(5961):92–94.

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

Papadopoulos C, Callebaut I, Gelly JC, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. bioRxiv. 2021; .

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

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

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

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. Origins of de novo genes in human and chimpanzee. PLoS genetics. 2015; 11(12):e1005721.

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

Saxer G, Havlak P, Fox SA, Quance MA, Gupta S, Fofanov Y, Strassmann JE, Queller DC. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba Dictyostelium discoideum. Plos One. 2012; .

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

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.

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.

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

Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines CP, Iannotta J, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nature communications. 2020; 11(1):1–18.

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.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. A molecular portrait of de novo genes in yeasts. Molecular Biology and Evolution. 2018; 35(3):631–645.

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

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

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.

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.

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

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.

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.

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

Fig.1–Figure Supplement 1. Estimates of one model time-step in various species. The following references were used to obtain mutation rates: ***Farlow et al.*** (***2015***); ***Keightley et al.*** (***2009***); ***Lee et al.*** (***2012***); ***Lind and Andersson*** (***2008***); ***Long et al.*** (***2016***); ***Ness et al.*** (***2012***); ***Ossowski et al.*** (***2010***); ***Saxer et al.*** (***2012***); ***Sung et al.*** (***2012***); ***Zhu et al.*** (***2014***). And the following references were used for generation times: ***Cassidy-Hanley*** (***2012***); ***Milo et al.*** (***2010***); ***Petersen and Russell*** (***2016***); ***Harris*** (***2009***); ***Ishida and Hori*** (***2017***); ***Fey et al.*** (***2007***); ***Silva et al.*** (***2009***); ***Fernández-Moreno et al.*** (***2007***); ***Koornneef and Scheres*** (***2001***). 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.

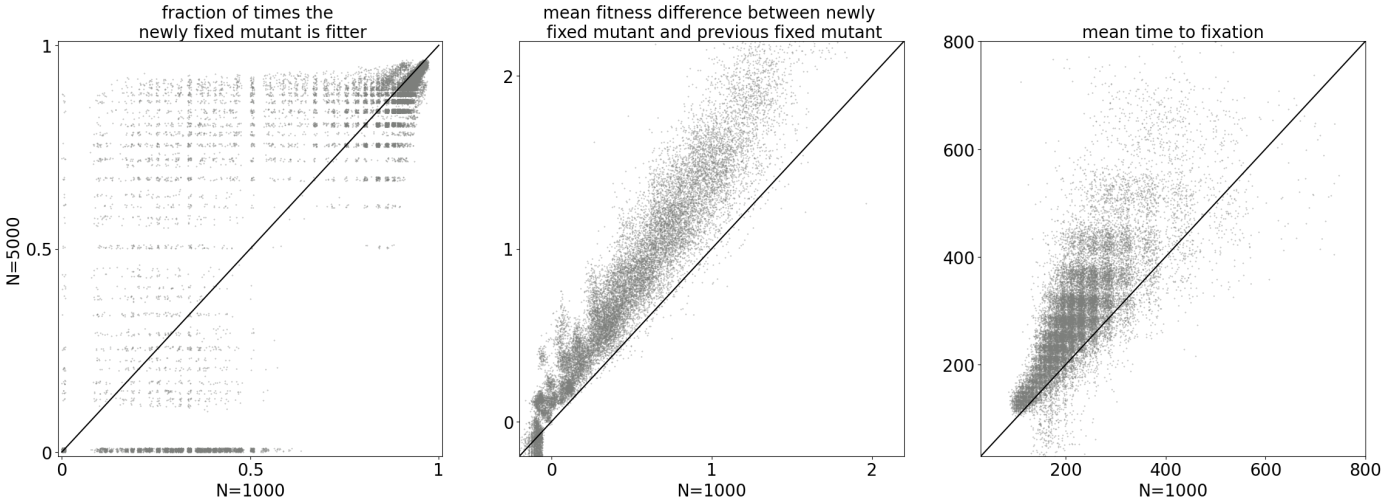

Fig.2–Figure Supplement 1. Scatter plots comparing fixation dynamics in N = 1000 and 5000 populations. For each population size, 22,500 systems were analysed across all parameter values. The x and y value for each point represents (A) Fraction of times fixation of fitter mutant occurred, (B) mean fitness 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

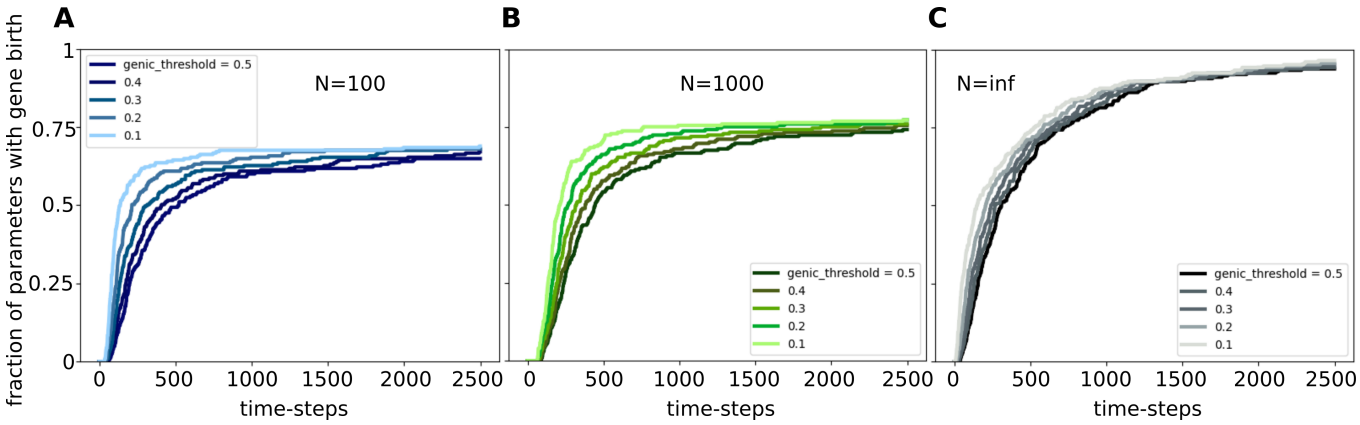

Fig.3–Figure Supplement 1. In (A,B,C) the x-axis represents time-steps and the y-axis represents the fraction of parameters tested for which gene birth occurred at least in 1 of the 100 replicate populations tested. Population sizes: (A) N=100, (B) N=1000, (C) infinite population.

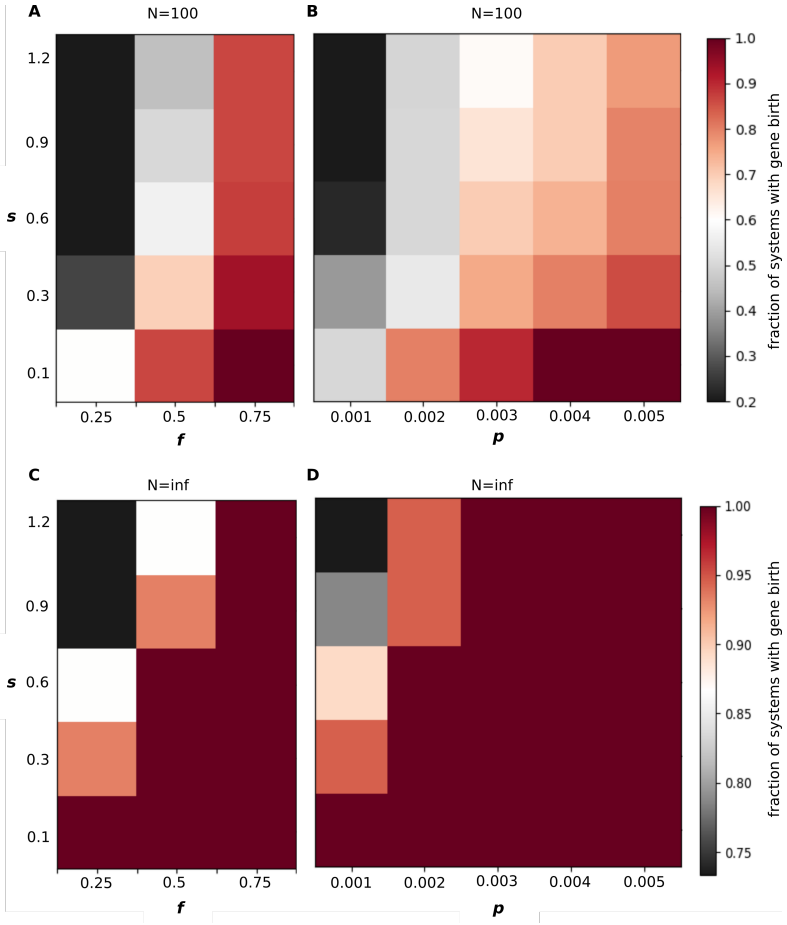

Fig.3–Figure Supplement 2. Trade-off between the shape parameter and frequency and size of beneficial mutations. We show results here for populations of sizes (A,B) N = 100 and (C,D) N = inf. In the heatmaps of rows indicate values of the shape parameter, columns indicate values of (A,C) fraction of 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.

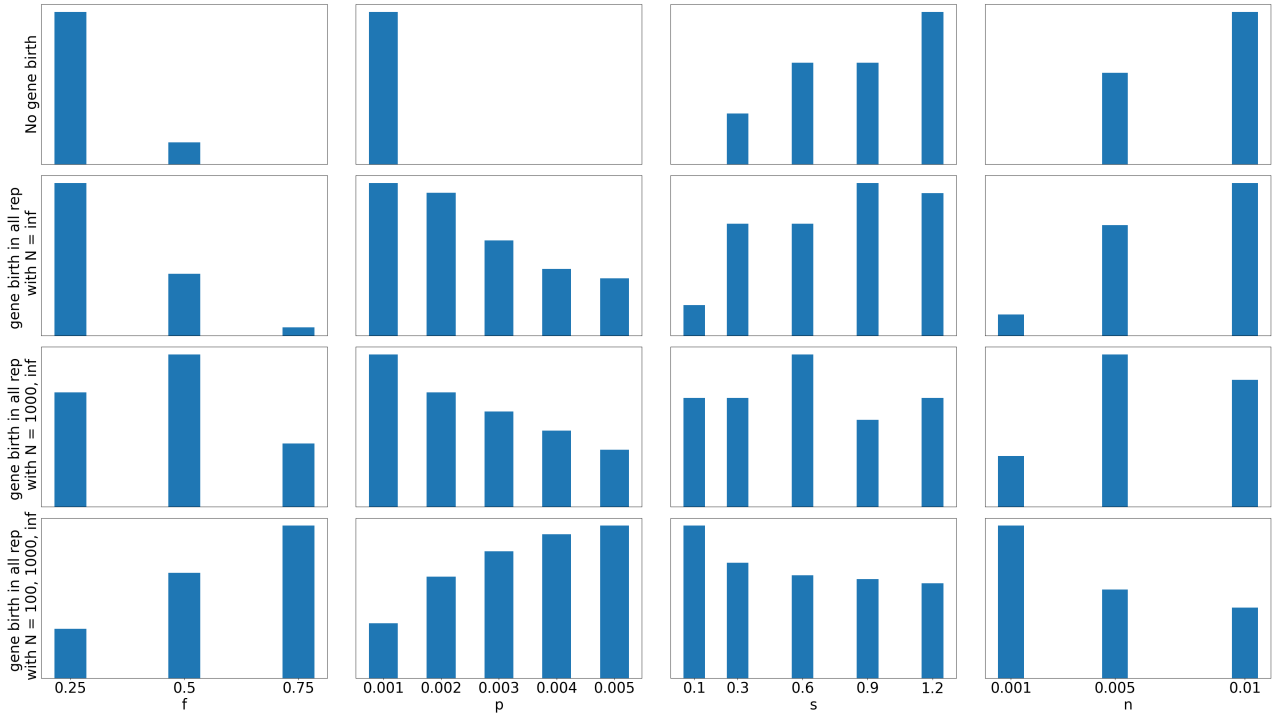

Fig.3–Figure Supplement 3. Effect of model parameters on gene birth probability. Each row contains histograms for parameter values at which *first row*: no gene birth occurred in any population, *second row*: gene birth occurred in all 100 replicates only in N = inf populations, *third row*: gene birth 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*

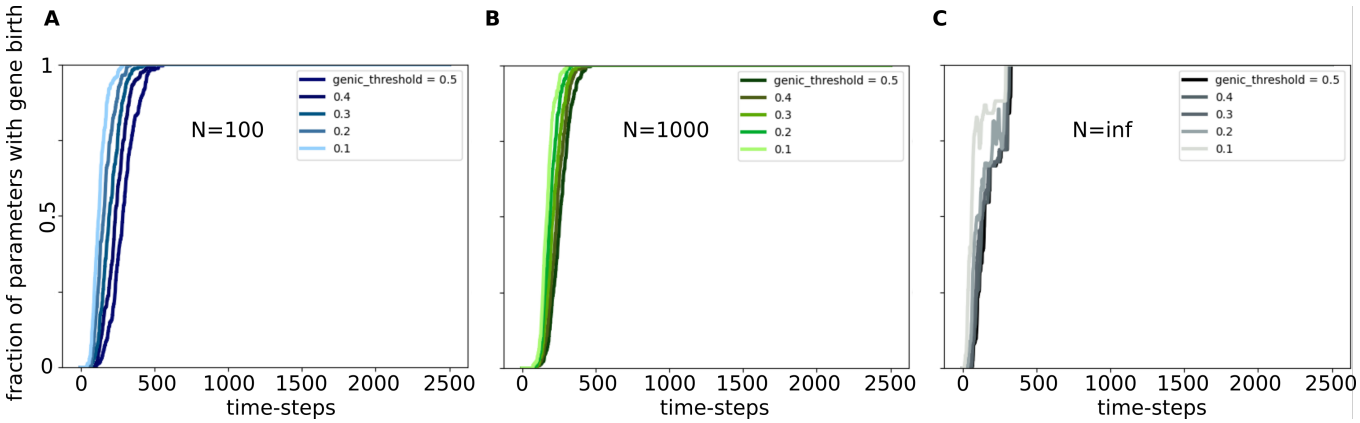

Fig.4–Figure Supplement 1. All parameters lead to gene birth within 2500 time-steps under the fluctuating DFE regime. In (A,B,C) the x-axis represents time-steps and the y-axis represents the fraction of parameters tested for which gene birth occurred at least in 1 of the 100 replicate populations tested. Population sizes: (A) N=100, (B) N=1000, (C) infinite population.

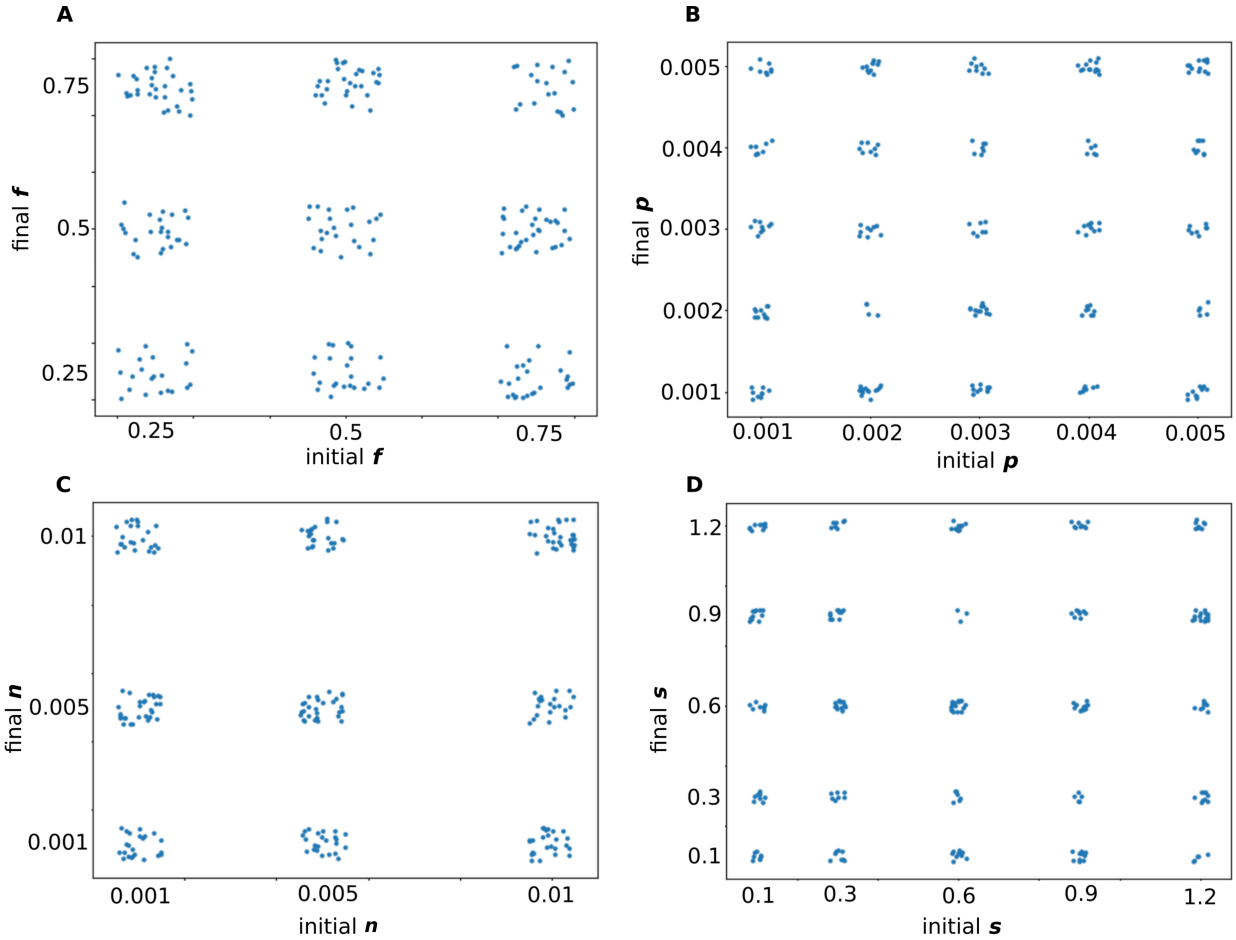

Fig.4–Figure Supplement 2. Scatter plots for initial and final values of parameters under fluctuating DFE regime. Each point represents a distinct N = 1000 population. 2250 systems were analysed for this figure. Noise has been added to make the density of points more apparent. (A) fraction of beneficial mutations *f*, (B) mean size of beneficial mutations *p*, (C) mean size of deleterious mutations *n*, (D) shape parameter *s*

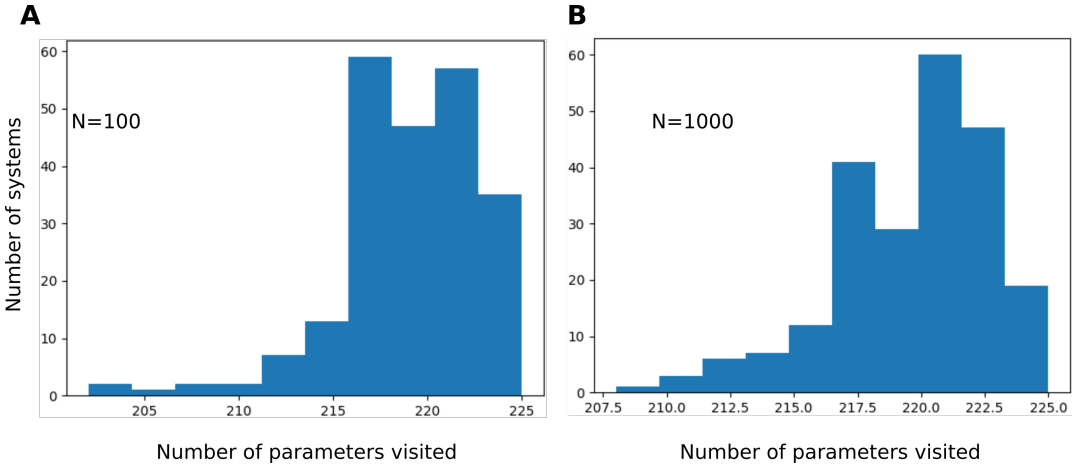

Fig.4–Figure Supplement 3. Histograms for number of distinct parameters visited by populations under the fluctuating DFE regime. 2250 systems were analysed for each histogram. (A) N = 100, (B) N = 1000