Wilson and Liberles
BMC Ecology and Evolution (2023) 23:14
https://doi.org/10.1186/s12862-023-02116-y

BMC Ecology and Evolution

RESEARCH

Open Access

Dosage balance acts as a time-dependent selective barrier to subfunctionalization

Amanda E. Wilson[1] and David A. Liberles[1*]

Abstract

Background Gene duplication is an important process for genome expansion, sometimes allowing for new gene functions to develop. Duplicate genes can be retained through multiple processes, either for intermediate periods of time through processes such as dosage balance, or over extended periods of time through processes such as subfunctionalization and neofunctionalization.

Results Here, we built upon an existing subfunctionalization Markov model by incorporating dosage balance to describe the interplay between subfunctionalization and dosage balance to explore selective pressures on duplicate copies. Our model incorporates dosage balance using a biophysical framework that penalizes the fitness of genetic states with stoichiometrically imbalanced proteins. These imbalanced states cause increased concentrations of exposed hydrophobic surface areas, which cause deleterious mis-interactions. We draw comparison between our Subfunctionalization + Dosage-Balance Model (Sub + Dos) and the previous Subfunctionalization-Only (Sub-Only) Model. This comparison includes how the retention probabilities change over time, dependent upon the effective population size and the selective cost associated with spurious interaction of dosage-imbalanced partners. We show comparison between Sub-Only and Sub + Dos models for both whole-genome duplication and small-scale duplication events.

Conclusion These comparisons show that following whole-genome duplication, dosage balance serves as a time-dependent selective barrier to the subfunctionalization process, by causing an overall delay but ultimately leading to a larger portion of the genome retained through subfunctionalization. This higher percentage of the genome that is ultimately retained is caused by the alternative competing process, nonfunctionalization, being selectively blocked to a greater extent. In small-scale duplication, the reverse pattern is seen, where dosage balance drives faster rates of subfunctionalization, but ultimately leads to a smaller portion of the genome retained as duplicates. This faster rate of subfunctionalization is because the dosage balance of interacting gene products is negatively affected immediately after duplication and the loss of a duplicate restores the stoichiometric balance. Our findings provide support that the subfunctionalization of genes that are susceptible to dosage balance effects, such as proteins involved in complexes, is not a purely neutral process. With stronger selection against stoichiometrically imbalanced gene partners, the rates of subfunctionalization and nonfunctionalization slow; however, this ultimately leads to a greater proportion of subfunctionalized gene pairs.

Keywords Whole genome duplication, Subfunctionalization, Dosage balance, Stochastic process model, Biophysical model, Genome content

*Correspondence:
David A. Liberles
daliberles@temple.edu
Full list of author information is available at the end of the article

Background

Gene duplication is a very important process in the evolution of genomes [1]. Many genes have undergone some type of duplication event in their history. The human lineage together with other vertebrates had two rounds of whole-genome duplication during the chordate-vertebrate transition [2, 3]. Plant phylogenies indicate many rounds of duplication events in their history [4–11]. Gene redundancy relaxes selection and makes faster evolutionary exploration of sequence space and gene function space possible [1, 12]. Duplicate copies may lead to the development of beneficial fitness effects, including novel morphological traits [13, 14]. To understand the evolutionary processes that follow gene duplication is to understand a source of genome expansion, pathway complexity, and functional innovation.

The process of gene duplication has been well studied in the past half century with two general categories of duplication events are small-scale duplication (SSD) events and large-scale events [15]. Small-scale duplication events affect relatively small sections of the genome. These errors usually occur during DNA replication or through transposition events [16]. Larger-scale duplication events include chromosomal or whole-genome duplication (WGD). These typically originate from nondisjunction events in meiosis and/or hybridization [17], and are typically rarer than their small-scale counterpart but can be beneficial [18, 19].

Genes that originate from whole-genome duplication and small-scale duplication events follow different patterns of retention [4, 20–24]. These different patterns occur because of their initial effect on the genome and cell function. Specifically, duplications and subsequent processes may affect the stoichiometric balance of gene products. Protein subunits often have a hydrophobic surface that are buried in a protein complex and function to aid in the binding of the complex. However, when subunits are not bound in their prospective protein complex, these hydrophobic surface areas are solvent exposed, so they will seek out a hydrophobic environment to bury into. The interactions caused by this force may lead to deleterious effects. Because duplicate copies often lead to higher expression, these copies can influence the stoichiometry of protein-complex subunits. Therefore, selection acts to maintain the stochiometric balance by removing expression of redundant copies to avoid these mis-interactions and aggregation of the gene products [25, 26].

Small-scale duplication events happen frequently, but they immediately interfere with the stoichiometric balance, so they tend not to be highly favorable, so selection favors the loss of these duplicated genes [23, 27, 28]. Large scale duplication events are more likely to include genes and their interacting partners, so there is selective pressure to keep the additional copies to avoid negatively affecting the stoichiometric balance [20, 29, 30]. Gene dosage balance describes the tendency for duplicated gene copies to be retained for an intermediate time immediately following the whole-genome duplication event [31–33]. This is thought to be why there is an initially fast gene loss rate of duplicate copies after small-scale duplication events, but a slower initial gene loss rate in larger duplication events, followed by a faster loss rate once the stoichiometric balance is affected [30, 33–42], unless this is counterbalanced by functional changes that affect fitness [33, 43, 44]. Fernández et al. [45] have also noted that beyond the increased waiting time for subfunctionalization enabled by dosage balance, that the two processes can interplay to affect fitness.

Nonfunctionalization, the process of making one of the copies of a gene no longer functional, is the most common fate of duplicated genes [1]. This is because there is a lot of opportunity to gain an early stop codon or some other loss of function mutation (for example, nonsense, frameshift), and the other copy can continue the ancestral function [39]. Of the genes that are ultimately retained for a long time, it is likely these extra copies have some sort of benefit or that they subfunctionalized; otherwise, probability dictates that random function degrading mutations would take over in a few million years [15, 46], either through neutral or selective processes. For duplicated copies that have generated a fitness advantage, there are a few generally accepted ways these gene copies can be beneficial. First, there is a small chance that it is beneficial to have extra copies of the same functionally exact gene as a way of upregulating that gene [47, 48], but that is likely not true for the vast majority of retained genes [49]. Other processes include regulatory neofunctionalization, coding function neofunctionalization, or specialization after subfunctionalization [1, 34, 50–52]. For genes that first subfunctionalize, it is also believed that subfunctionalization can be an intermediate step to ultimate neofunctionalization, in a process called subneofunctionalization [53, 54]. This is supported by the observation that duplicate genes that are retained over long evolutionary time periods do show patterns consistent with both neofunctionalization and subfunctionalization [55].

An important research question in the field is, "what makes some genes have large numbers of duplicates while other genes exclusively exist as singletons?". One of the leading theories to explain this phenomenon is the gene duplicability hypothesis, that certain types of genes, whether that "type" is categorized by GO terms or by the complexity of its interaction network/pathway, are more likely to benefit from extra copies of these genes [56, 57]. This benefit may be because they may be more likely to gain new functions or it is favorable and mutationally accessible to specialize [46, 56, 58, 59]. Additionally, dosage balance may play a role in retention of the extra copies, especially depending on what kind of duplication event has occurred in genome history and when. Duplicated genes in dosage balance can also have effects on the expression of and interactions with other genes and their protein products across the genome. Such trans-acting effects can contribute to the selective landscape of the evolution of individual gene duplicate pairs in parallel to changes in the gene itself and its regulatory regions [60–64]

As previously suggested, as a whole-genome duplication event ages and is under the influence of gene dosage balance, the probability of retaining duplicate gene copies decreases [33, 34, 46, 57, 65]. Therefore, gene dosage balance is one important reason why retaining or losing duplicate gene copies after a whole-genome duplication event is a time-heterogeneous process. However, the dynamics and constraints of this process have not been fully explored. Previous attempts at modeling the effect that gene dosage balance has on duplicate gene copy retention were less mechanistic, utilizing survival analysis described by an increasing hazard function of duplicated and redundant gene copies [33, 34]. This type of model is a mathematical description of the observed phenomenon, it does not actually model the underlying biological process itself at the level of detail described here and therefore would not enable discovery of the effects of the process on the dynamics that are not obvious from data fitting.

Here, building upon an existing framework for subfunctionalization proposed by Force and Lynch [46] and developed as a full model by Stark et al. [59], we propose an alternative time-homogeneous model. This new model joins a chemical thermodynamic model with a population model to explore the selective pressures on the retention of duplicated genes through subfunctionalization when genes are influenced by dosage balance effects. Our model incorporates an element of fitness that depends on the stoichiometric balance. This fitness parameter models selective effects on gene duplicate retention through subfunctionalization. Subfunctionalization is typically conceptualized as a neutral process because each necessary function continues to be performed by one of the copies. We explored how losing functional expression of one of the copies affects the stoichiometric balance between that gene product and the other gene products. We expect that if a gene is sensitive to dosage balance effects, losing expression of one of the duplicate gene copies in a specific tissue or developmental stage will negatively affect the stoichiometric balance of gene products in that tissue or developmental stage, causing selection to act against both subfunctionalization and nonfunctionalization. We expect selection to act against any process that negatively affects the stoichiometric balance of gene products, including subfunctionalization because it would cause imbalance in specific expression domains, but even more so for nonfunctionalization because it affects the balance in all expression domains simultaneously. We built a modeling framework that produces this behavior by modeling the underlying process.

Methods

Calculating the sum concentration of exposed hydrophobic residues across expression domains

To ultimately model dosage balance, we are interested in estimating the magnitudinal effect that stoichiometric imbalance has on the fitness of each state of duplicate gene pairs. We model a heterodimer, with subunits A and B, which are transcribed and translated from Gene A (G_A) and Gene B (G_B) respectively. The heterodimer's binding interface is formed by one binding site on each subunit, and the binding site consists of a patch of exposed hydrophobic residues. We expect there to be some concentration of unbound subunit A ($[A]_{free}$), unbound subunit B ($[B]_{free}$), as well as subunits A and B in their bound form as a heterodimer ([AB]) in a cell. The reaction that yields the bound form can be seen in Eq. 1. All parameter and variable symbols, definitions, and values are described in Table 2.

$$A + B \leftrightarrow AB \tag{1}$$

The equilibrium constant (K_{eq}) for the reaction in Eq. 1 is based upon free energy differences. Quantitative data that is measurable for this kind of reaction include the K_{eq}, the total concentration of the A subunit from the expression of G_A ($[A]_{total}$), and total concentration of the B subunit from the expression of G_B ($[B]_{total}$). We can use this quantifiable information to calculate $[A]_{free}$, $[B]_{free}$, [AB] and ultimately the concentration of patches of exposed hydrophobic residues ([hp]) (Eqs. 2, 3, and 4). The total concentration of a subunit is the sum of the concentration of subunits in the unbound form and the concentration of subunits in the bound form (Eqs. 3 and 4).

$$[hp] = [A]_{free} + [B]_{free} \tag{2}$$

$$[A]_{total} = [A]_{free} + [AB] \tag{3}$$

$$[B]_{total} = [B]_{free} + [AB] \tag{4}$$

We can solve for the concentration of patches of exposed hydrophobic residues ([hp]) based on the above information by plugging in the [AB] solution above into the quadradic equation to get $[\mathrm{A}]_{\mathrm{free}}$ and $[\mathrm{B}]_{\mathrm{free}}$ in terms of K_{eq}, $[\mathrm{A}]_{\mathrm{total}}$, and $[\mathrm{B}]_{\mathrm{total}}$, which are known (Eq. 5). Once we can solve for [AB] we can calculate [hp] for each regulatory domain.

$$[\mathrm{AB}] = \frac{(K_{\mathrm{eq}}[\mathrm{A}]_{\mathrm{total}} + K_{\mathrm{eq}}[\mathrm{B}]_{\mathrm{total}} + 1) \pm \sqrt{(-(K_{\mathrm{eq}}[\mathrm{A}]_{\mathrm{total}} + K_{\mathrm{eq}}[\mathrm{B}]_{\mathrm{total}} + 1)^2 - 4K_{\mathrm{eq}}(K_{\mathrm{eq}} * [\mathrm{A}]_{\mathrm{total}} * [\mathrm{B}]_{\mathrm{total}}))}}{2K_{\mathrm{eq}}} \quad (5)$$

The conceptualize the stoichiometric imbalance "load" to be the sum of [hp] across each regulatory domain. Therefore, we assume the fitness of each state is inversely proportional to the sum of hydrophobic patches across z regulatory domains. Because of this relationship, we calculate fitness (*f*) using an inverse function, with the relationship between the sum of hydrophobic patches per cell and the corresponding fitness penalty scaled by w (Eq. 6).

$$f = \frac{1}{(1 + w\sum_{1\to z}[\mathrm{hp}]_z)} \quad (6)$$

Loss of expression of duplicate copies causes stoichiometric imbalance. The imbalance of gene products introduces a fitness consequence. We want to use see how the fitness consequences affects the probability of fixing such a loss mutation in a population. We use an existing framework to calculate the probability of fixing mutations (*g*) [66], plugging Eq. 6 into the fitness terms, with the relative fitness of the current state being f_i and the relative fitness of the next possible state being f_j and N_e being the effective population size (Eq. 7). We will use this calculation to calculate the rates between states in our Markov chain in the next section.

$$g = \frac{1 - \frac{f_i}{f_j}}{1 - \frac{f_i}{f_j}^{N_e}} = \frac{1 - \frac{1+(w\sum_{1\to z}[\mathrm{hp}]_z)_j}{1+(w\sum_{1\to z}[\mathrm{hp}]_z)_i}}{1 - \frac{1+(w\sum_{1\to z}[\mathrm{hp}]_z)_j}{1+(w\sum_{1\to z}[\mathrm{hp}]_z)_i}^{N_e}} \quad (7)$$

Continuous-time Markov chain

We define a continuous-time Markov chain $\{X(t), t \geq 0\}$ $($Figure1a$)$ for our Subfunctionalization + Dosage (Sub + Dos) Model that is similar to that of the Stark et al. [59] model we refer to as, the Subfunctionalization-Only (Sub-Only) Model (Fig. 1b). We use the same state space,

$$\mathcal{A} = \{0, 1, \ldots, z-1\} \cup \{S, Y\}, \quad (8)$$

and state $i \in \{0, 1, \ldots, z-1\}$ each represent a set of duplicate gene pairs where one duplicate copy has that number of nonfunctional regulatory regions. State S represent a duplicate gene pair that has been subfunctionalized and State Y represent a duplicate gene pair where one copy has been lost (pseudogenized). Both State S and Y are absorbing states.

Like that of the Sub-Only Model [59], our Sub + Dos Model is also based on the mechanics of regulatory subfunctionalization, the assumption that knock-out mutations occur at a constant rate and are independent of each other, and that selection ensures that at least one copy of each regulatory region is retained. Also, like Stark et al. [59] model, we assume a haploid genome to avoid the complication of recombination. Results would be at least partially readily extendable to the diploid case with a natural model for the role of dominance resulting from the explicit link with underlying biochemistry. This makes the model more readily extendable from haploid to diploid cases than purely statistical genetics models that are naïve to the underlying biochemical processes that generate patterns of dominance. The model does not deal with recombination between alleles or context-dependent allele behavior in a diploid setting. Duplicates are assumed to be fixed in the genome. It might be noted that during the early phases of duplication in a haploid setting, the duplicates can behave like alleles in a diploid population.

In contrast with the Sub-Only Model [59], our Sub + Dos Model incorporates dosage balance affects by assuming that the process of losing a regulatory function is non-neutral and that the probability of fixing a loss is not simply $1/N_e$. Instead, we assume that fitness is inversely proportional to the magnitude of dosage imbalance introduced by the loss. We estimate dosage imbalance stoichiometrically through the concentration of exposed hydrophobic residues ([hp]) (Eq. 6). Therefore, the relative fitness of each state of duplicate gene pairs is inversely proportional to the sum of [hp] across expression domains.

We use the fitness of each state, that incorporates the fitness penalty associated with loss of expression, to calculate the probability of fixation [66] (Eq. 7) and use that to calculate the rate of transition between states (equation set 11). Because the fitness of each state is only affected by the sum of [hp], the probability of fixing a loss mutation is therefore determined by the concentration of exposed hydrophobic residues that loss introduces. We chose to model the effect of dosage balance in this way because is consistent with expected underlying mechanisms [26, 67].

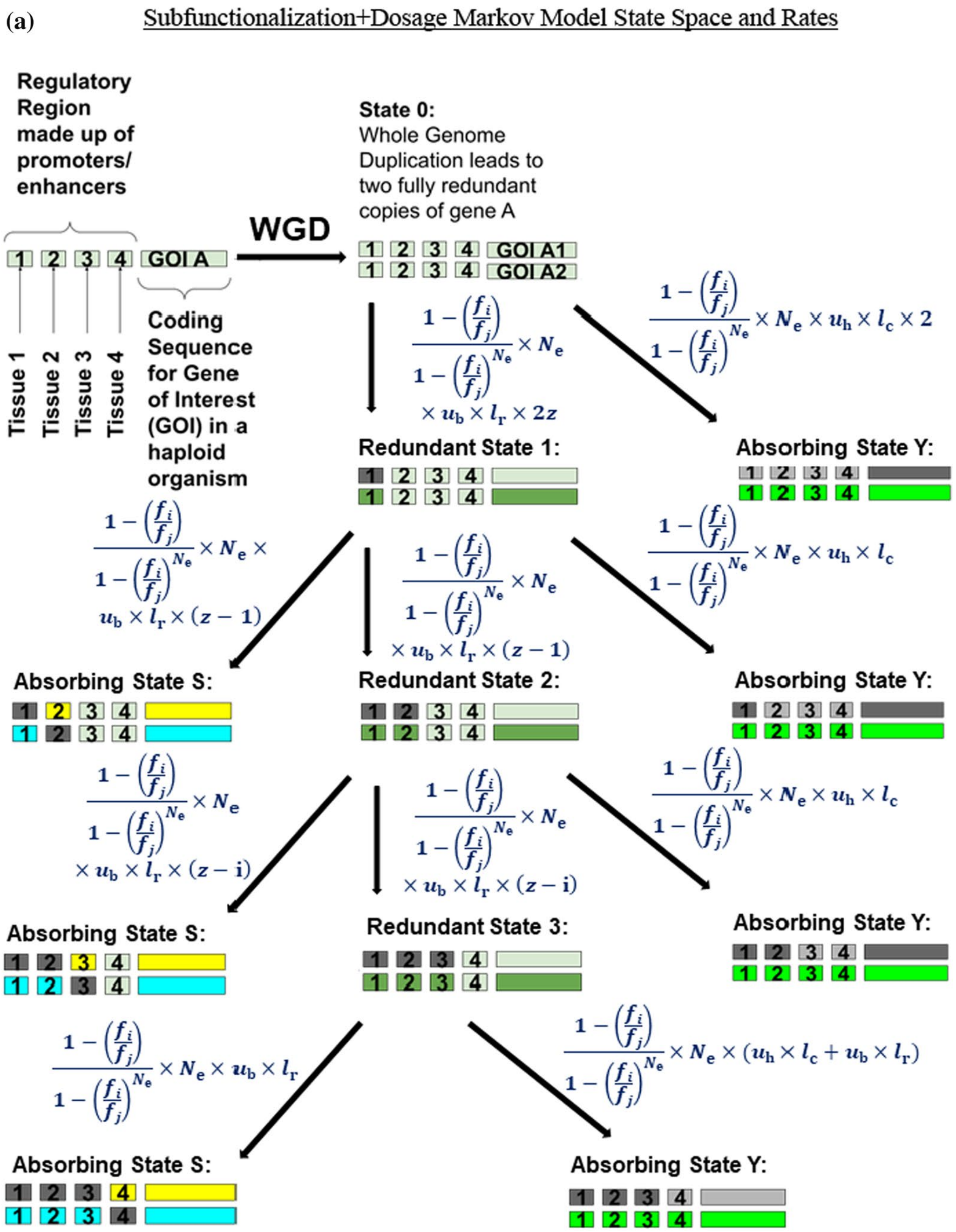
(a)
Subfunctionalization+Dosage Markov Model State Space and Rates
Regulatory Region made up of promoters/ enhancers
1 2 3 4 GOI A
Tissue 1
Tissue 2
Tissue 3
Tissue 4
Coding Sequence for Gene of Interest (GOI) in a haploid organism
WGD
State 0: Whole Genome Duplication leads to two fully redundant copies of gene A
1 2 3 4 GOI A1
1 2 3 4 GOI A2
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r \times 2z$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_h \times l_c \times 2$
Redundant State 1:
Absorbing State Y:
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r \times (z-1)$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r \times (z-1)$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_h \times l_c$
Absorbing State S:
Redundant State 2:
Absorbing State Y:
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r \times (z-i)$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r \times (z-i)$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_h \times l_c$
Absorbing State S:
Redundant State 3:
Absorbing State Y:
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times u_b \times l_r$
$\frac{1-\left(\frac{f_i}{f_j}\right)}{1-\left(\frac{f_i}{f_j}\right)^{N_e}} \times N_e \times (u_h \times l_c + u_b \times l_r)$
Absorbing State S:
Absorbing State Y:
1 2 3 4

Fig. 1 Markov model for the fate of retention of duplicate copies after gene duplication. In this example, the gene of interest (GOI A) has been duplicated into two copies, GOI A_1 and GOI A_2. Both copies of GOI A have been duplicated with all four of their regulatory domains upstream of the gene. Each regulatory domain acts as an enhancer for tissues 1–4. The state space includes the two copies with full redundancy (State 0), transient unresolved states (States 1–3), and absorbing states (States Y, S) indicated by neon colors. The absorbing states include nonfunctionalization of one of the gene copies (State Y) and subfunctionalization that leads to the retention of both copies (State S). The neon colors represent which copy is permanently retained and what function(s) it preforms. The light green parts indicate unresolved portions of the gene. The parts of the gene that are dark grey indicate that part being knocked-out through mutation. The parts of the gene that are light grey indicate parts of the gene that are no longer functional because of mutations that occurred in other parts of the gene. Note that for both part a and b, the rate equation for State 1 → State S is equal to the rate equation for State 1 → State 2 AND the rate equation for State 2 → State S is equal to the rate equation for State 2 → State 3. **a** Subfunctionalization + Dosage Model. The formulas for the rates between states are calculated using a fixation probability equation [66], which uses the relative fitness of the current state (f_i) and the next state (f_j), and the effective population size (N_e). The rates also incorporate the number of regulatory domains (z), the nucleotide length of the regulatory domains (l_r), the nucleotide length of the coding region (l_c), the loss of function nucleotide mutation rate (u_b, u_h). **b** Subfunctionalization-Only Model [59] for the fate of retention of duplicate copies after gene duplication. The formulas for the rates between states are calculated the effective population size (N_e), the number of regulatory domains (z), Poisson rate at which null mutations are fixed in each of the z mutable regulatory regions for each gene (u_r, Eq. 10), and the Poisson rate at which null mutations fix in the coding regions (u_c, Eq. 9)

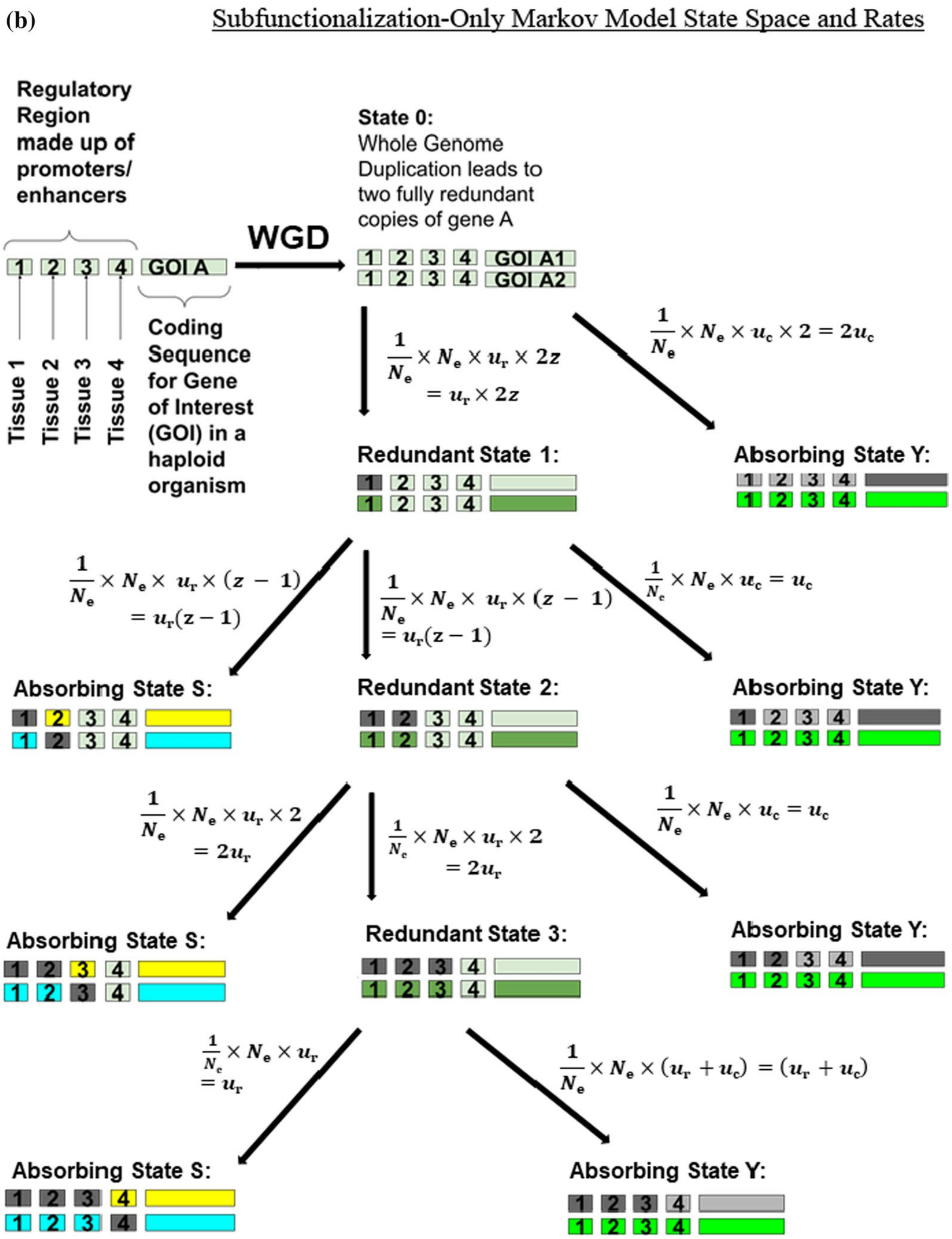
(b)
Subfunctionalization-Only Markov Model State Space and Rates
Regulatory Region made up of promoters/ enhancers
1 2 3 4 GOI A
Tissue 1
Tissue 2
Tissue 3
Tissue 4
Coding Sequence for Gene of Interest (GOI) in a haploid organism
WGD
State 0: Whole Genome Duplication leads to two fully redundant copies of gene A
1 2 3 4 GOI A1
1 2 3 4 GOI A2
$\frac{1}{N_e} \times N_e \times u_r \times 2z = u_r \times 2z$
$\frac{1}{N_e} \times N_e \times u_c \times 2 = 2u_c$
Redundant State 1:
Absorbing State Y:
$\frac{1}{N_e} \times N_e \times u_r \times (z-1) = u_r(z-1)$
$\frac{1}{N_e} \times N_e \times u_r \times (z-1) = u_r(z-1)$
$\frac{1}{N_c} \times N_e \times u_c = u_c$
Absorbing State S:
Redundant State 2:
Absorbing State Y:
$\frac{1}{N_e} \times N_e \times u_r \times 2 = 2u_r$
$\frac{1}{N_c} \times N_e \times u_r \times 2 = 2u_r$
$\frac{1}{N_e} \times N_e \times u_c = u_c$
Absorbing State S:
Redundant State 3:
Absorbing State Y:
$\frac{1}{N_c} \times N_e \times u_r = u_r$
$\frac{1}{N_e} \times N_e \times (u_r + u_c) = (u_r + u_c)$
Absorbing State S:
Absorbing State Y:
1 2 3 4

Fig. 1 continued

Additionally, we expanded the parameters set from those in Stark et al. [59] for the Sub-Only model to better reflect empirical data. We conceptualize the Poisson rate at which null mutations fix in the coding regions (u_c) and the Poisson rate at which null mutations are fixed in each of the z mutable regulatory regions for each gene (u_r) as being calculated by specific types of mutations that lead to the loss of function and the opportunity for those mutations that can be empirically measured. We calculate the rate of loss of a regulatory domain (u_r) as the product of the nucleotide rate of certain mutations that impair transcription factor binding (u_b) and the nucleotide length of regulatory regions (l_r) (Eq. 10). We calculate the rate of loss of the coding region (u_c) as the product of nucleotide rate of mutations that lead to a non-functional mRNA or peptide chain (u_h) and the nucleotide length of

Table 1 The Q Matrix form for the new Subfunctionalization + Dosage Model, given in equation set 10

	0	**1**	**2**	**3**	**S**	**Y**
0	$-\frac{1-\left(\frac{f_0}{f_1}\right)}{1-\left(\frac{f_0}{f_1}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot 2z$ $-2 * \frac{1-\left(\frac{f_0}{f_Y}\right)}{1-\left(\frac{f_0}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$	$\frac{1-\left(\frac{f_0}{f_1}\right)}{1-\left(\frac{f_0}{f_1}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot 2z$	0	0	0	$2 * \frac{1-\left(\frac{f_0}{f_Y}\right)}{1-\left(\frac{f_0}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$
1	0	$-\frac{1-\left(\frac{f_1}{f_2}\right)}{1-\left(\frac{f_1}{f_2}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-1)$ $-\frac{1-\left(\frac{f_1}{f_S}\right)}{1-\left(\frac{f_1}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-1)$ $-\frac{1-\left(\frac{f_1}{f_Y}\right)}{1-\left(\frac{f_1}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$	$\frac{1-\left(\frac{f_1}{f_2}\right)}{1-\left(\frac{f_1}{f_2}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-1)$	0	$\frac{1-\left(\frac{f_1}{f_S}\right)}{1-\left(\frac{f_1}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-1)$	$\frac{1-\left(\frac{f_1}{f_Y}\right)}{1-\left(\frac{f_1}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$
2	0	0	$-\frac{1-\left(\frac{f_2}{f_3}\right)}{1-\left(\frac{f_2}{f_3}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-2)$ $-\frac{1-\left(\frac{f_2}{f_S}\right)}{1-\left(\frac{f_2}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-2)$ $-\frac{1-\left(\frac{f_2}{f_Y}\right)}{1-\left(\frac{f_2}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$	$\frac{1-\left(\frac{f_2}{f_3}\right)}{1-\left(\frac{f_2}{f_3}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-2)$	$\frac{1-\left(\frac{f_2}{f_S}\right)}{1-\left(\frac{f_2}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-2)$	$\frac{1-\left(\frac{f_2}{f_Y}\right)}{1-\left(\frac{f_2}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$
3	0	0	0	$-\frac{1-\left(\frac{f_3}{f_S}\right)}{1-\left(\frac{f_3}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-3)$ $-\frac{1-\left(\frac{f_3}{f_Y}\right)}{1-\left(\frac{f_3}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$ $-\frac{1-\left(\frac{f_3}{f_Y}\right)}{1-\left(\frac{f_3}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r$	$\frac{1-\left(\frac{f_3}{f_S}\right)}{1-\left(\frac{f_3}{f_S}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r \cdot (z-3)$	$\frac{1-\left(\frac{f_3}{f_Y}\right)}{1-\left(\frac{f_3}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_h \cdot l_c$ $+\frac{1-\left(\frac{f_3}{f_Y}\right)}{1-\left(\frac{f_3}{f_Y}\right)^{Ne}} \cdot N_e \cdot u_b \cdot l_r$
S	0	0	0	0	0	0
Y	0	0	0	0	0	0

Model parameters and variables are defined in Table 2. Equations to calculate the fitness parameter (*f*) can be found in Eq. 6

coding region (l_c) (Eq. 9). However, in this paper we do not explore the effects of changing the rates of mutations that impair transcription factor binding and those that lead to a nonfunctional protein for simplicity. Instead, we make the reasonable assumption that both rates of loss of function would be caused by similar types of mutations, including but not limited to mutations that directly affect transcription factor binding and mutations that affect the phasing of DNA binding sites, therefore having comparable nucleotide mutation rates. However, in application of our model, these rates can easily be different from those employed here.

$$u_{\mathrm{c}} = (u_{\mathrm{h}} \cdot l_{\mathrm{c}} \cdot N_{\mathrm{e}})/N_{\mathrm{e}} = u_{\mathrm{h}} \cdot l_{\mathrm{c}} \tag{9}$$

$$u_{\mathrm{r}} = (u_{\mathrm{b}} \cdot l_{\mathrm{r}} \cdot N_{\mathrm{e}})/N_{\mathrm{e}} = u_{\mathrm{b}} \cdot l_{\mathrm{r}} \tag{10}$$

Therefore, the generator matrix for the Subfunctionalization + Dosage Markov Chain is defined to be Q = $[q_{ij}]$, where the matrix form of Q is shown in Table 1. The nonzero off-diagonals are given by q_{ij} (equation set 11). A defense for these transition rates can be found in Stark et al. [59] in conjunction with our above argument on the changes. Because Q is the generator/transition rate matrix, the rows sum to 0. To accomplish this, all other i,j^{th} entries are 0 except the diagonals, which are zero minus the sum off the defined row terms. Again, where S and Y are absorbing states with S referring to the Subfunctionalization State, and Y referring to the Pseudogenization/Nonfunctionalization state.

$$q_{ij} = \begin{cases} 2 * g_{0,Y} * N_{\mathrm{e}} * u_{\mathrm{h}} * l_{\mathrm{c}}, & \textit{if} \;\; i = 0, j = Y \\ 2 * z * g_{0,1} * N_{\mathrm{e}} * u_b * l_{\mathrm{r}}, & \textit{if } i = 0, j = 1 \\ g_{i,Y} * N_{\mathrm{e}} * u_{\mathrm{h}} * l_{\mathrm{c}}, & \textit{if } 1 \le i \le z-2, j = Y \\ (z-i) * g_{i,j} * N_{\mathrm{e}} * u_{\mathrm{b}} * l_{\mathrm{r}}, & \textit{if } 1 \le i \le z-2, j = i+1 \\ (z-i) * g_{i,S} * N_{\mathrm{e}} * u_{\mathrm{b}} * l_{\mathrm{r}}, & \textit{if} 1 \le i \le z-2, j = S \\ g_{z-1,Y} * N_{\mathrm{e}} * u_{\mathrm{b}} * l_{\mathrm{r}} + g * N_{\mathrm{e}} * \; u_{\mathrm{h}} * l_{\mathrm{c}}, & \textit{if } \; i = z-1, j = Y \\ g_{z-1,Y} * N_{\mathrm{e}} * u_{\mathrm{b}} * l_{\mathrm{r}}, & \textit{if } \; i = z-1, j = S. \end{cases} \tag{11}$$

We define the probability matrix as P = $[p_{ij}]$. For each generation, P is calculated by exponentiating e to the product of time in the number of generations and Q.

Probability distribution calculations

The computer program, written in C++, can be found at https://github.com/aewilson96/Wilson_Liberles_2022, calculates the probability distribution of states for a pair of genes for each generation following a whole-genome duplication event. Because of the nature of Markov Chains, we can directly calculate the probability distribution from the generator matrix; therefore, there is no need to run simulations over time. The calculation for the rate of transitioning from one state to the other, includes the probability of fixation (g, Eq. 7), the effective population size (N_e), the nucleotide rate of loss of function mutations (u_b and u_h), the length of the coding sequence (l_c), and the length of the regulatory domain (l_r, enhancer, promoter, silencer). The probability of fixation calculation [66] utilizes the relative fitness of the current state (f_i) and the relative fitness of the next possible state (f_j). To determine the relative fitness of each state, our method assumes an inverse relationship between the summation of the concentration of exposed hydrophobic patches summed across expression domains and fitness (Eq. 6). Using the equilibrium constant (K_{eq}), the concentration of hydrophobic residues ([hp]) is calculated for when the total concentration of gene A products and gene B products are in stoichiometric balance, and for when they are in a 1:2 ratio, which is the expected imbalance for a pair of gene homologs if one copy is not functionally expressed (Eq. 6). Then, for each state, these values are summed across expression domains, being the summation of both (1) the product of unaffected domains and the concentration of hydrophobic patches when they are in stoichiometric balance and (2) the product of the affected domains and the concentration of hydrophobic patches when they are stoichiometrically imbalanced.

Small-scale duplication events work a little differently than whole-genome duplication events. These events cause immediate stochiometric imbalance and losing a copy of the duplicated gene is expected to repair the balance. Because of this difference, the concentration of hydrophobic residues is expected to be greater immediately after the duplication event, and nonfunctionalization/pseudogenization becomes the state with the highest fitness because it repairs the stoichiometric balance quickest, therefore having the highest probability of fixing. Additionally, losing redundancy and subfunctionalizing are still more favorable states than the totally redundant State 0. This is different than in the whole-genome duplication case, because for whole-genome duplication events, losing expression in anyway yields a lower fitness (because of the higher concentration of hydrophobic residues); therefore, these dosage effects slow progression through the states towards the absorbing states. In small-scale duplication events, this progression is faster when dosage balance effects act.

All of the figures that show the Subfunctionalization + Dosage (Sub + Dos) Model have 4 regulatory regions/domains (z), a scalar of 1.0 (w), a nucleotide mutation rate that affects transcription of a functional coding strand of 2.5×10^{-8} nucleotide mutations per generation (u_h), a nucleotide mutation rate that affects transcription binding to regulatory region/domain of 2.5×10^{-8} nucleotide mutations per generation (u_b),

Table 2 The list of symbols and definitions for parameters and variables, which are used and/or defined in Eqs. 1–11

Parameters and Variables	Symbol	Realistic range based on empirical data	Range used in provided figures
Poisson rate of nucleotide mutations that interfere with transcription binding	u_b	1.0×10^{-8}–2.5×10^{-8} nucleotide mutations per generation [70]	2.5×10^{-8} nucleotide mutations per generation
Poisson rate of nucleotide mutations that interfere with production of a functional peptide sequence	u_h	1.0×10^{-8}–2.5×10^{-8} nucleotide mutations per generation [70]	2.5×10^{-8} nucleotide mutations per generation
Length of coding region	l_c	5.0×10^{4} nucleotides [71]	5.0×10^{4} nucleotides
Length of enhancer	l_r	50 bp to 1.5 kbp [72]	775 nucleotides
Subfunctionalization-Only Model poisson rate at which null mutations fix in the coding regions [59]	u_c	–	Equation 9, loss of function mutations per generation
Subfunctionalization-Only Model poisson rate at which null mutations are fixed in each of the *z* mutable regulatory regions for each gene [59]	u_r	–	Equation 10, loss of function mutations per generation
Number of regulatory regions/domains (promoters/enhancers/silencers)	z	< 20 regulatory regions	4 regulatory regions
Effective population size	N_e	1.4×10^{4} individuals	1×10^{2}–1.4×10^{6} individuals
Time in generations since duplication event	t	5.0×10^{3}–1.4×10^{4}generations	5.0×10^{3}–1.0×10^{4}generations
Equilibrium constant	K_{eq}	1.0×10^{6}–1.0×10^{14} mol/mL [73, 74]	1.0×10^{4}–1.0×10^{12} mol/mL
Scalar on the relationship between the number of hydrophobic patches per cell and the corresponding fitness penalty	w	–	1.0
Heterodimer of interest	AB	–	Equation 1
Gene that codes for subunit A in heterodimer AB	$\mathrm{G_A}$	–	–
Gene that codes for subunit B in heterodimer AB	$\mathrm{G_B}$	–	–
Subunit of heterodimer of interest, gene product of $\mathrm{G_A}$	A	–	Equation 1
Subunit of heterodimer of interest, gene product of $\mathrm{G_B}$	B	–	Equation 1
Concentration of the heterodimer of interest that is in its bound form	[AB]	–	
Total concentration of subunit A, likely to be estimated by transcription data	$[\mathrm{A}]_{\mathrm{total}}$	1×10^{-6}–1×10^{-10} mol/mL [75, 76]	2.5×10^{-6} mol/mL
Total concentration of subunit B, gene product of $\mathrm{G_B}$, likely to be estimated by transcription data	$[\mathrm{B}]_{\mathrm{total}}$	1×10^{-6}–1×10^{-10} mol/mL [75, 76]	2.5×10^{-6} mol/mL
Concentration of subunit A in the unbound form	$[\mathrm{A}]_{\mathrm{free}}$	–	Equation 3, in mol/mL
Concentration of subunit B in the unbound form	$[\mathrm{B}]_{\mathrm{free}}$	–	Equation 4, in mol/mL
Concentration of exposed hydrophobic patches	[hp]	–	Equation 2, in mol/mL
Current state	i	–	–
Next possible state	j	–	–
Fitness of state	f	–	Equation 6
Fitness of current state	f_i	1.0	Equation 7
Fitness of next possible state	f_j	–	Equation 7
Probability of fixation of mutation	g	–	Equation 7 [66]
Subfunctionalization, an absorbing state	S	–	Equations 8 and 11
Pseudogenization/nonfunctionalization, an absorbing state	Y	–	Equations 8 and 11

Potential ranges for each parameter are listed, as well as the range used in the production of the figures provided

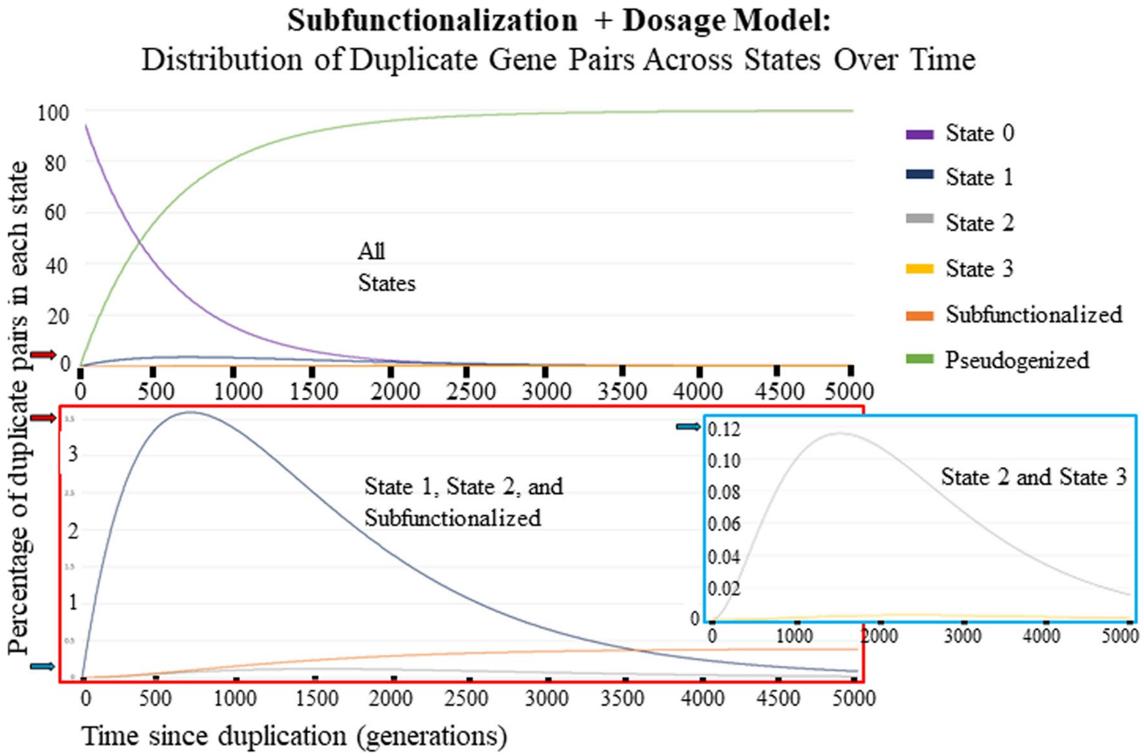

Fig. 2 Distribution of Duplicate Gene Pairs across States over 5000 generations, represented as a percentage. The purple line is the percentage of gene pairs that are completely redundant. The dark blue line is the percentage of gene pairs in State 1 with one of the copies having lost one expression domain. The gray line is the percentage of gene pairs in State 2 with one of the copies having lost two expression domains. The yellow line is the percentage of gene pairs in State 2 with one of the copies having lost three of four expression domains. The orange line is the percentage of gene pairs where the expression domains have been subfunctionalized. The green line is the percentage of gene pairs where one of the copies have been completely nonfunctionalized/pseudogenized. The red box is a zoomed in graph of the percentage of gene pairs in State 1, State 2 and Subfunctionalized. The blue box is a zoomed in graph of the percentage of gene pairs in State 2 and State 3. The red arrow indicates where 3.5% is on the y axis. The blue arrow indicates where 0.12% is on the y axis

5.0×10^4 nucleotide long coding regions, and 775 nucleotide long regulatory region/domains. The equivalent parameters used for all figures that show the Subfunctionalization-Only (Sub-Only)[59] Model is 4 regulatory regions/domains (z), 1.25×10^{-3} mutations per generation that affects transcription binding to regulatory region/domain($u_r = u_b \cdot l_c$), 1.9375×10^{-5} mutations per generation that affect transcription of a functional coding strand ($u_c = u_h \cdot l_r$). Figures 2, 3, 4, and 5 all used a K_{eq} value of 1.0×10^{10} mol/mL. Figure 6 used a range of K_{eq} values including 1.0×10^4; 1.0×10^6; 1.0×10^9; 1.0×10^{12} (mol/mL). For all figures, the concentration of total A was 2.5×10^{-6} mol/mL immediately after duplication. The same is true for concentration of total B, except for Fig. 3b, because that models a small-scale duplication event, where B was not duplicated so the concentration of total B used was 1.25×10^{-6} mol/mL. The figures show anywhere between 2.0×10^3 and 1.0×10^4 generations after the duplication event, chosen based on figure clarity. The effective population size chose for each figure was 1.4×10^5 individuals for Figs. 2 and 3. For Figs. 4, 5a–e and 6, we used a range of N_e's from as low as 100 to as much as 1.0×10^7.

All computer software is available on Github at https://github.com/aewilson96/Wilson_Liberles_2022.

Expected impact of parameter choices

We wanted to use parameter values that existed in a realistic range in order to show the magnitude of the effect that dosage balance has on the rate of subfunctionalization. To obtain biologically realistic values, we conducted a literature search for values that were equal to our parameters, or a proxy that would be on a similar order of magnitude as our parameters (see Tables 2 and 3). We chose these values to perform our calculations because they represented close proxies and had good evidence for them; however, it is expected that the general behavior shown in the results section holds true regardless of the values used. The logical argument for this expectation is as follows.

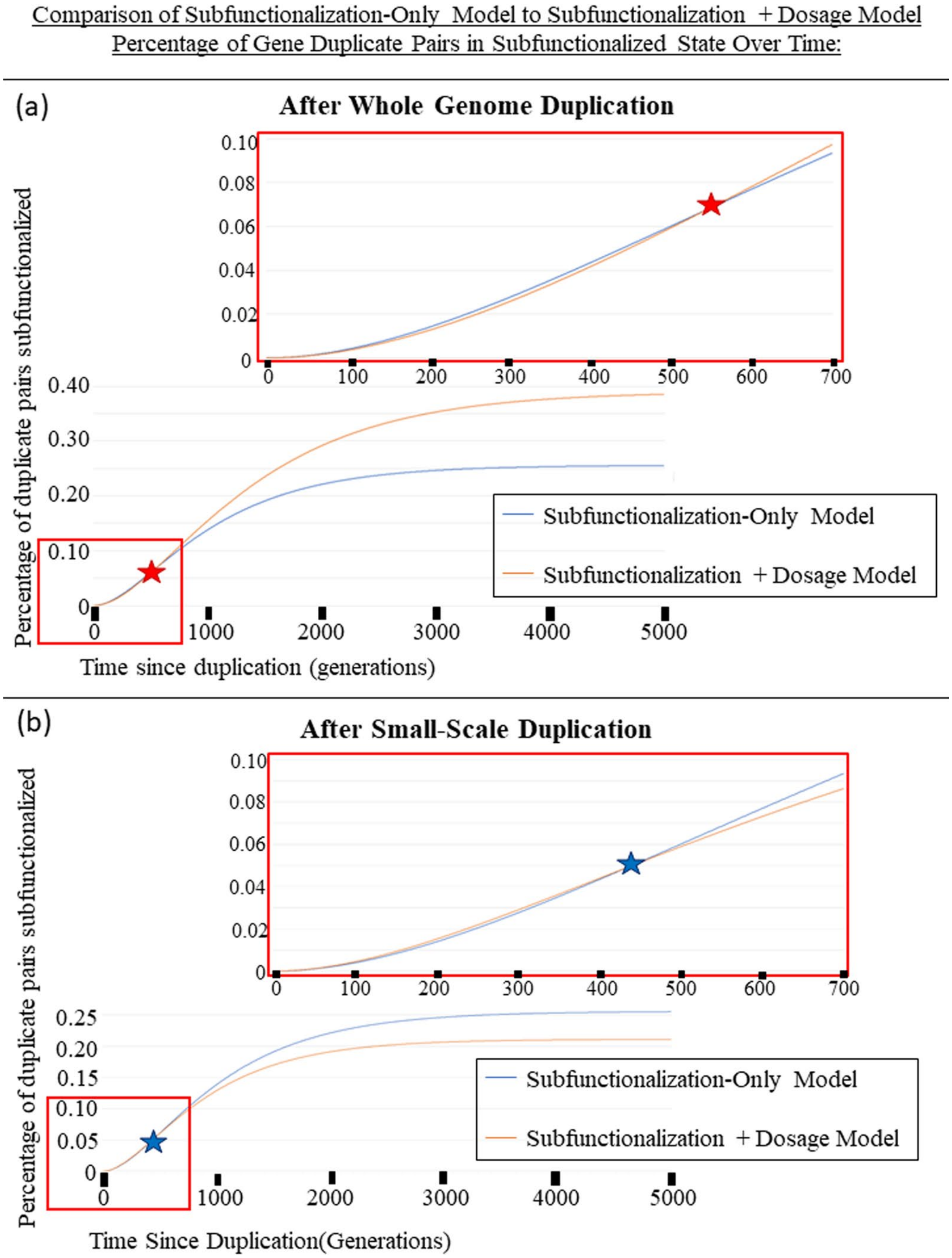

Fig. 3 **a** The percentage of gene pairs that have been subfunctionalized over 5000 generations after a Whole-genome Duplication Event. The blue line is the Subfunctionalization-Only Model. The orange line is the new Subfunctionalization + Dosage Model. The red box shows the graph zoomed in to 700 generations and 0.1% gene pairs. The red star represents where the two lines cross, prior to the star, Sub-Only Model has a higher percentage of gene pairs that are subfunctionalized, while after the star, the Sub + Dos model has a higher percentage of gene pairs that have been subfunctionalized. **b** The percentage of gene pairs that have been subfunctionalized over 5000 generations after a Small-Scale Duplication Event. The blue line is the Subfunctionalization-Only Model. The orange line is the new Subfunctionalization + Dosage Model. The red box shows the graph zoomed in to 700 generations and 0.1% gene pairs. The blue star represents where the two lines cross, prior to the star, Sub-Only Model has a lower percentage of gene pairs that are subfunctionalized, while after the star, the Sub + Dos model has a lower percentage of gene pairs that have been subfunctionalized

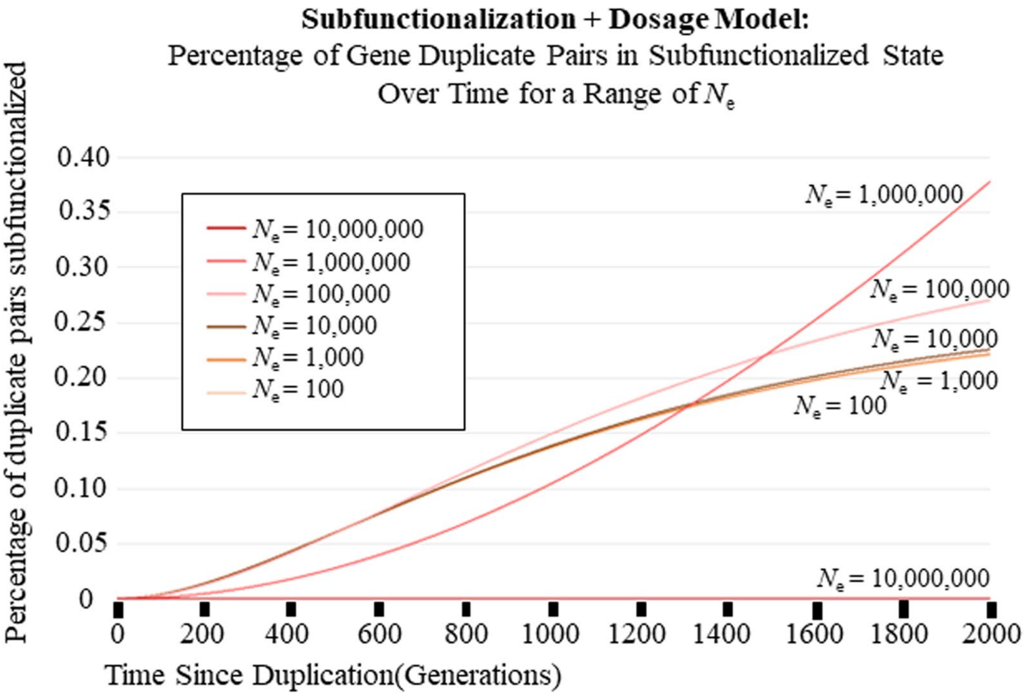

Fig. 4 The percentage of gene pairs that have been subfunctionalized over 2000 generations after a whole-genome Duplication Event for 6 different effective population sizes (N_e) for the new Sub + Dos Model. Note that as the effective population size increases, so does the efficacy of selection, and that leads to the pattern where there is a longer delay for subfunctionalization to occur, but will ultimately lead to a higher percentage of subfunctionalized duplicate gene pairs. Also note that with the largest N_e shown is so delayed, that it hasn't even begun to subfunctionalize for the number of generations shown, however will ultimately surpass the others in the percentage of subfunctionalized gene pairs

The z values are equal for the rate calculations for both the Subfunctionalization-Only Model [59]and Subfunctionalization + Dosage Model. As shown in Eqs. 9 and 10, u_c is equivalent to $u_h \cdot l_c$, and u_r is equivalent to $u_b \cdot l_r$. Additionally, u_h, l_c, u_b, and $l_r > 0$, because the mutation rate would always be positive, and the length of the regions will be positive. Therefore, u_c and $u_r > 0$.

So, we can ignore u_c and u_r from Sub-Only Model and $u_h \cdot l_c$ and $u_b \cdot l_r$ from Sub + Dos Model because they are equal. The only difference in the rate calculation of the Sub + Dos Model from the Sub-Only Model is $g \cdot N_e$ equals 1 in the Sub-Only equation (because $N_e \cdot 1/N_e = 1$), but Sub + Dos calculates g as the probability of fixation from the relative fitnesses of each state [66].

(See figure on next page.)

Fig. 5 A comparison of the effect that N_e has on the resulting percentage of gene pairs that have subfunctionalized after a whole-genome duplication event for the Subfunctionalization-Only Model and the Subfunctionalization + Dosage Model. The blue lines are the Subfunctionalization-Only Model and the orange lines are the Subfunctionalization + Dosage Model (**a**) Subfunctionalization-Only Model over 7000 generations for 3 different effective population sizes (N_e). Note that the percentage of subfunctionalized genes is the same at any given time because the effective population size does not affect the rate of subfunctionalization in this model. **b** Over 7000 generations for 3 different effective population sizes (N_e) for our Subfunctionalization + Dosage Model. Note that as the effective population size increases, so does the efficacy of selection, and that leads to the pattern where there is a longer delay for subfunctionalization to occur, but will ultimately lead to a higher percentage of subfunctionalized duplicate gene pairs. **c** Over 6000 generations comparing the Subfunctionalization-Only Model to our Subfunctionalization + Dosage Model with effective population sizes (N_e) = 1000. The red star represents where the two lines cross, prior to the star, Sub-Only Model has a higher percentage of gene pairs that are subfunctionalized, while after the star, our Sub + Dos model has a higher percentage of gene pairs that have been subfunctionalized. **d** Over 6000 generations comparing the Subfunctionalization-Only Model to our Subfunctionalization + Dosage Model with effective population sizes (N_e) = 100,000. The red star represents where the two lines cross, prior to the star, Sub-Only Model has a higher percentage of gene pairs that are subfunctionalized, while after the star, our Sub + Dos model has a higher percentage of gene pairs that have been subfunctionalized. **e** Over 5000 generations comparing the Subfunctionalization-Only Model to our Subfunctionalization + Dosage Model for two effective population sizes (N_e) = 1000 (lighter colored lines) and N_e = 100,000 (darker colored lines). For each N_e value, initially the Sub-Only model has a higher percentage of gene pairs that are subfunctionalized, but eventually our Sub + Dos Model has a higher percentage of gene pairs that have been subfunctionalized

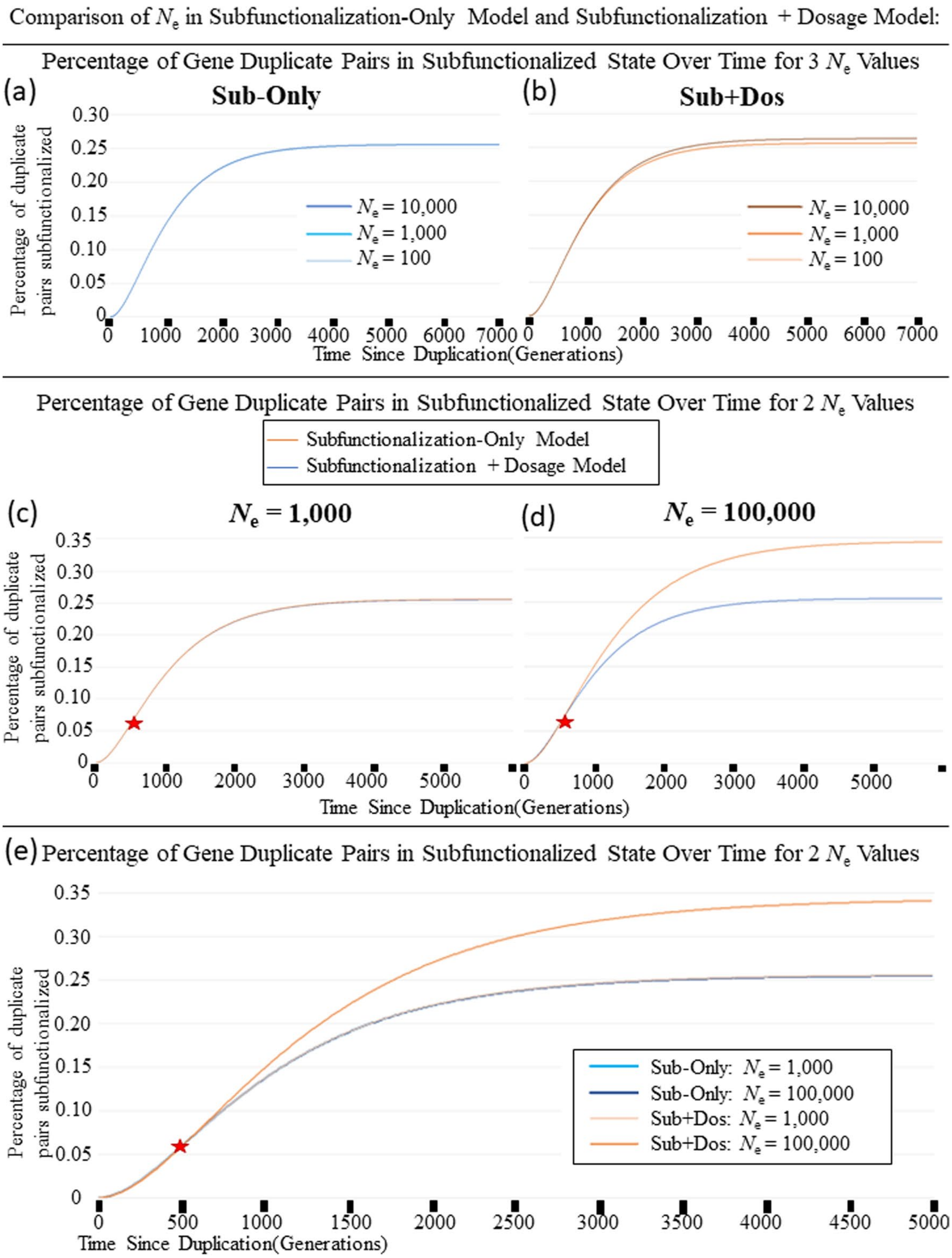

Fig. 5 (See legend on previous page.)

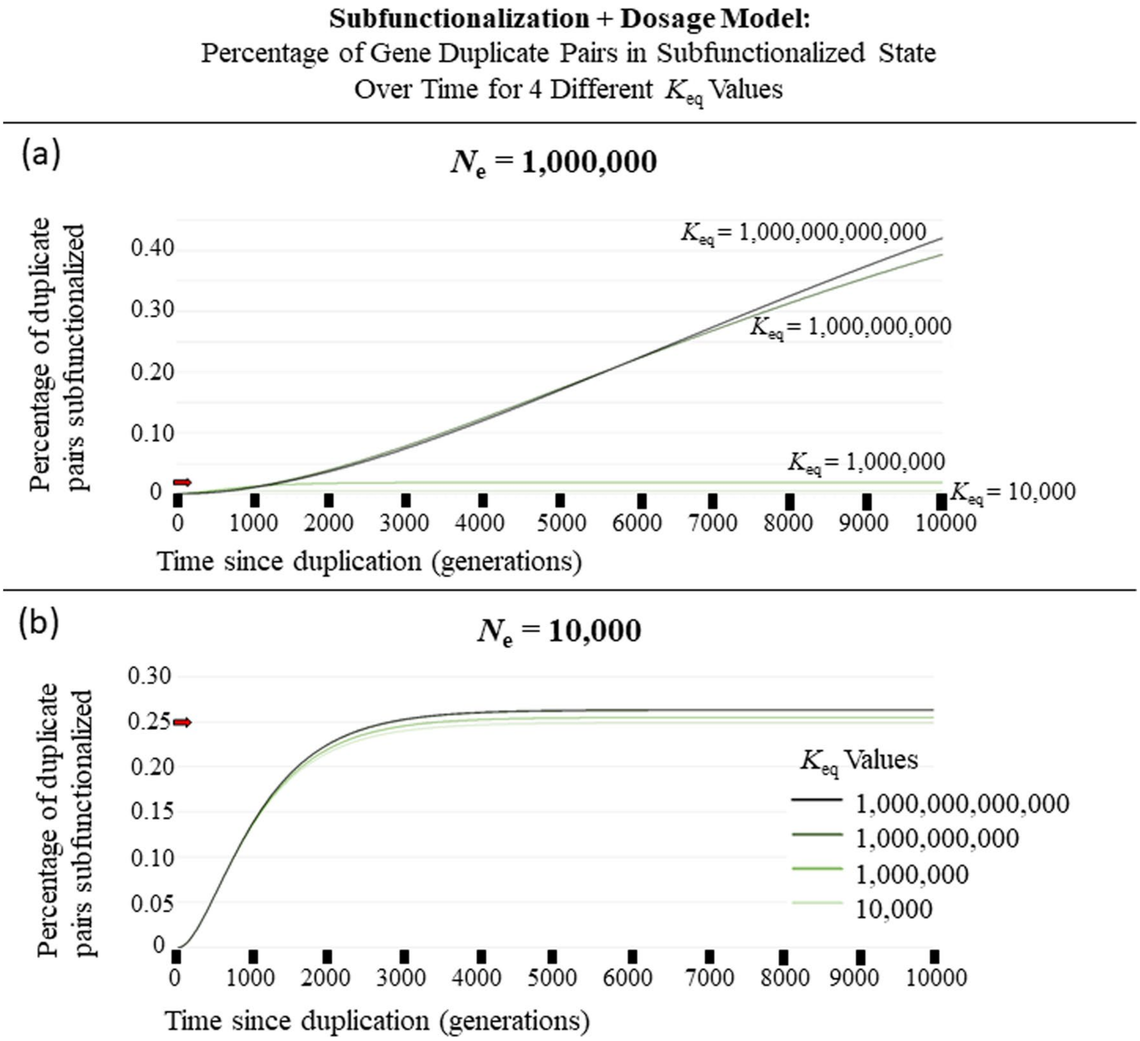

Fig. 6 The percentage of gene pairs that have been subfunctionalized over 10,000 generations with an effective population size (N_e) of **a** 1,000,000 **b** 10,000 after a Whole-genome Duplication Event for 4 different equilibrium constants (K_{eq}) for our Subfunctionalization + Dosage Model. Note that as K_{eq} increases, so does the strength of selection, and that leads to the pattern where there is a longer delay for subfunctionalization to occur but will ultimately lead to a higher percentage of subfunctionalized duplicate gene pairs

So therefore, the important factor is how the behavior of rate = 1 is different than $g \cdot N_e$.

Because in whole-genome duplication events, [hp] will increase to some extent with the introduction of more imbalance, even if it is fractionally small, the fitness of the next state will always be lower than that of the current state. A number larger than 1 to the N_e^{th} power, given the N_e^{th} power is larger than 1 (which is a reasonable assumption for population sizes), will also be larger than 1. Therefore, the rate of transitioning to the next state will always be less than 1 for the Sub + Dos Model, given these assumptions, so it will always have a smaller rate than the rate in the Sub-Only Model. Note that having more introduced imbalance will always make the fitness of that state lower relatively, so nonfunctionalization will be less favorable than subfunctionalization for $z \geq 2$, so the rate of nonfunctionalization will be even lower than the rate of subfunctionalization in the Sub + Dos Model.

Therefore, we should expect to see the same pattern presented by our results after a whole-genome duplication events for parameters that abide by our assumptions. To summarize, these assumptions include that the nucleotide mutation rate is positive, the length of the nucleotide regions is positive, subfunctionalization and nonfunctionalization are neutral in the Sub-Only model, subfunctionalization and nonfunctionalization receive a fitness penalty associated with the extent of imbalance the state introduces, and the population size is greater than 1.

Table 3 Parameters that change for the Subfunctionalization + Dosage Model for each figure

Figure number	K_{eq} (mol/mL)	N_e (individuals)	Time (generations)	$[A]_{total}$ immediately after duplication (mol/mL)	$[B]_{total}$ immediately after duplication (mol/mL)
2	10,000,000,000	140,000	5000	0.0000025	0.0000025
3a	10,000,000,000	140,000	5000 (zoomed in 700)	0.0000025	0.0000025
3b (Small-Scale)	10,000,000,000	140,000	5000 (zoomed in 700)	0.0000025	0.00000125
4	10,000,000,000	100; 1,000; 10,000; 100,000; 1,000,000; 10,000,000	2000	0.0000025	0.0000025
5a (Sub-Only)	10,000,000,000	100; 1,000; 10,000	7000	0.0000025	0.0000025
5b (Sub + Dos)	10,000,000,000	100; 1,000; 10,000	7000	0.0000025	0.0000025
5c	10,000,000,000	100,000; 1,000	6000	0.0000025	0.0000025
5d	10,000,000,000	100,000; 1,000	6000	0.0000025	0.0000025
5e	10,000,000,000	1,000; 100,000	5000	0.0000025	0.0000025
6	10,000; 1,000,000; 1,000,000,000; 1,000,000,000,000	10,000; 1,000,000	10,000	0.0000025	0.0000025
	10,000,000,000	100; 1,000; 10,000; 100,000; 1,000,000; 10,000,000	2000	0.0000025	0.0000025

The Subfunctionalization-Only Model shares the same time parameter for the corresponding figure

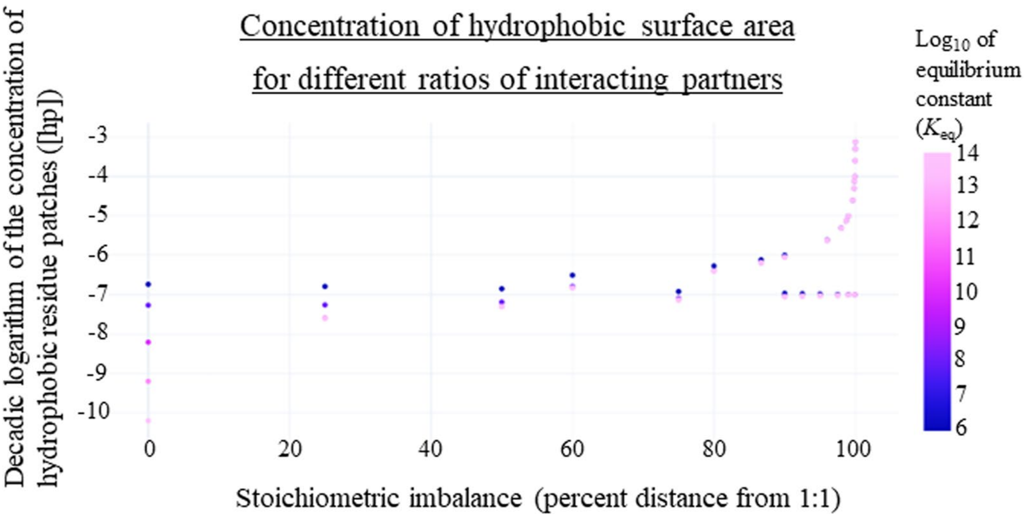

Fig. 7 Concentration of hydrophobic surface area for different ratios of interacting partners. Concentration of hydrophobic surface area is shown on a log scale, as well as equilibrium constant values (K_{eq}). When the K_{eq} values are larger, and the ratio of interacting partners ($[A]_{total}/[B]_{total}$) is closer to one, the sum of $[A]_{free}$ and $[B]_{free}$ is smaller. Similarly, when the K_{eq} values are smaller, and the ratio of interacting partners ($[A]_{total}/[B]_{total}$) is further from one, the sum of $[A]_{free}$ and $[B]_{free}$ is larger. This is possibly explained by the fact that overall exposed hydrophobic patches would be lower if A and B are in balance and are tight binders

Results

Figure 1a shows the state space and transition rates for the Subfunctionalization + Dosage-Balance Model. Figure 1b shows the Subfunctionalization-Only Model taken from Stark et al. [59], which has the same state space, but different transition rates. There are several similarities between the transition rate calculations for these two models. One similarity is that they both have the same opportunity for a new mutation to occur (mutation rates) because the state space is the same. The main difference is how the probability of fixation is calculated. The Sub-Only model combines several parameters into

Table 4 Comparison of the resulting percentage of genes in each state for Sub-Only Model and Sub + Dos Model is shown below for 400 Generations and 4,000,000 Generations following whole-genome duplication

Percentage of genes in each state	400 Generations Sub-Only	4,000,000 Generations Sub-Only	400 Generations Sub + Dos	4,000,000 Generations Sub + Dos
State 1: fully redundant (transient)	34.58%	0%	52.19%	0%
State 2: Loss of 1 enhancer (transient)	2.805%	0%	3.283%	0%
State 3: Loss of 2 enhancers on one copy (transient)	0.03559%	0%	0.03486%	0%
State 4: Loss of 3 enhancers on one copy (transient)	0.0001927%	0%	0.0001603%	0%
State S: Subfunctionalized (absorbing)	0.04339%	0.2557%	0.04053%	0.4687%
State Y: One copy pseudogenized (absorbing)	62.54%	99.74%	44.46%	99.53%

one parameter, with the probability of fixing a mutation (u_r or u_c) as being the product of the allele frequency in the population ($1/N_e$), the population size (N_e), the rate of nucleotide mutation rate (u_b or u_h), and the length of the region in question (l_r or l_c). Because there is one allele per individual as both models assume haploidy, $1/N_e$ and N_e cancel out and are therefore not included in their rate calculations. Because of the cancelation of parameters, the Sub-Only Model's u_r and u_c can be directly compared to the product of the nucleotide mutation rate (u_b or u_h), and the length of the region in question (l_r or l_c) in the Sub + Dos Model. Therefore, for the purpose of comparison, we assumed their u_r to be equal to $u_b \cdot l_r$ and u_c to be equal to $u_h \cdot l_c$. This expansion of parameter space was necessary to properly mirror the complexity of the biological processes involved in dosage compensation and allows us to input empirical information into the rate equations. While the Sub + Dos is also a haploid model, it uses the relative fitnesses of each state to calculate the probability of fixation [66]; therefore these values are included in the rate calculations. These fitnesses are correlated with the total concentration of hydrophobic residues, which are a representation of the amount of imbalance each state introduces between interacting partners.

Our results show that with more imbalanced interacting partners, there is an increased concentration of cellular exposed hydrophobic patches, which would lead to more spurious deleterious reactions occurring (Fig. 7). In addition, we found that with smaller K_{eq}, we also see this increased concentration of exposed hydrophobic patches because of the lower affinity, leading to more gene products to exist in the unbound state, also leading to more spurious deleterious interactions (Fig. 7). Also notably, the more imbalanced the partners are, the bigger the difference is between [hp] values with a small versus large K_{eq}. For Fig. 7, stoichiometric imbalance is presented as a percent distance away from the 1:1 ratio, where the ratios were scaled where 0% was a 1:1 ratio and 100% imbalanced is the ratio with the largest magnitude, therefore the furthest from 1:1. If $[A]_{total} \geq [B]_{total}$, we used $[B]_{total}/[A]_{total}$ as the ratio and if $[B]_{total} > [A]_{total}$ then, we used $[A]_{total}/[B]_{total}$ as the ratio.

The parameters used for the Sub + Dos Model across all figures are 4 regulatory regions/domains (z), fitness scalar of 1.0 (w), 2.5×10^{-8} nucleotide mutations per generation (u_h and u_b), coding region 50,000 nucleotides long (l_c), and regulatory region 775 nucleotides long (l_r). The equivalent parameters used for the Sub-Only Model across all figures are 4 regulatory regions/domains (z), 1.25×10^{-3} coding region mutations per generation ($u_r = u_b \cdot l_c$), and 1.9375×10^{-5} regulatory region mutations per generation ($u_c = u_h \cdot l_r$). Table 3 lists the parameters that are not the same throughout all the figures. Figure 2 shows the percentage of gene pairs in each state over 5000 generations for the Sub + Dos Markov model using parameter values found in Table 1. The Markov model begins in a fully redundant state and two absorbing states are possible, nonfunctionalization and subfunctionalization. As expected, under this modeling framework using these parameters, > 99% of redundant genes nonfunctionalize and a small fraction subfunctionalize.

Figure 3a shows 5,000 generations after a whole-genome duplication event. When using equivalent parameters, the initial rate of subfunctionalization parameters is higher with the Sub-Only Model than in the new Sub + Dos Model after a whole-genome duplication event. However, as indicated by a star, a transition occurs and the equilibrium level of subfunctionalization is much higher with the Sub + Dos model. Table 4 shows the frequencies of different states over time under the different models.

Alternatively, Fig. 3b shows 5000 generations after a small-scale duplication event. It shows how the comparison of the behaviors between models would change after a small-scale duplication. Again, here we show that the initial subfunctionalization rate is faster with the

Sub + Dos model, but is eventually overtaken by the Sub-Only model, which is in fact, the opposite as it is after a whole-genome duplication event. This pattern is intuitive because dosage balance is initially preserved after a whole-genome duplication event, but not after a small-scale event, making the fitness of losing a mutation after a small-scale event more favorable.

The results obtained in Fig. 2 are dependent upon choices of parameters (Table 3). A range of parameter values exploring the effects of N_e (and thereby selection) (Fig. 5b–e, Additional file 1: Fig. S1) and of equilibrium binding constants (Fig. 6) were explored. Figure 4, shows how the Sub + Dos model changes for 6 different N_e values between 100 and 1.0×10^7 individuals over 2000 generations. These results show that as the N_e increases, so does the efficacy of selection, resulting in a longer delay for subfunctionalization to occur but eventually leading to more subfunctionalized gene pairs, and this holds true when compared to the Sub-Only model (Fig. 5b–e, Additional file 1: Fig. S1). This effect, a longer delay but ultimately more subfunctionalized pairs, was also observed for increasing binding affinity.

Discussion

When dosage balance effects are added to a model for subfunctionalization, this leads to increased retention after whole-genome duplication events after an initial delay in the rate of subfunctionalization. Consistent with observations on differential retention patterns for smaller-scale duplication, the opposite trends are observed, with a reduction in the probability of terminal subfunctionalization.

It should be noted that while most duplicate genes are lost and are lost relatively quickly from genomes, this process is slowed down in whole-genome duplication events relative to small-scale duplication events in genomic data [3, 30] and is slowed down for duplicates that are dosage balanced [55, 68, 69]. The model has tunable parameters for the selective strength that will affect the absolute levels of retention, but the qualitative effects are observable over broad ranges of parameterization.

Dosage balance as a process generates a time-dependent selective barrier to subfunctionalization and to nonfunctionalization. The dynamics of this process involve delayed terminal subfunctionalization, but subfunctionalization at higher rates in the end. Because this is selective and is dependent upon N_e, it emerges that subfunctionalization of genes when dosage balance processes are acting is not a purely neutral process. This is a finding that has not previously been described in the literature to our knowledge.

While we have not independently varied *w*, the selective scalar, this becomes convoluted with N_e in determining selective effects and variation in *w* would be expected to mirror variation in N_e. Variation in K_{eq} reveals that higher K_{eq} values, that favor subunits in their bound form, also leads to increased delay but higher rates of terminal subfunctionalization, because it increases the selection against imbalanced proteins because fewer subunits are in the unbound state when in balance.

In this study, we have assumed that gene expression levels of each duplicate remain constant. Ascencio et al. [49] found that in fact and as expected, gene expression evolves as a co-evolutionary process after gene duplication. While this was not modeled for the sake of simplicity, changing gene expression as a stochastic process could be added to the model to examine the dynamics under those scenarios. Further, the interactions that were modeled were those that reflected a heterodimer that forms a stable interaction. The extension to trimers and higher order heterocomplexes is possible, and is expected to yield similar results, but no longer enables simple analytical transition probabilities of the type described here. Additionally, the main findings are expected to hold true for any gene that is sensitive to dosage balance effects. Another simplifying assumption was the expression in only two tissues. Stark et al. [59] explored the role of the number of tissues expressed in the dynamics of subfunctionalization without dosage; that complexity could also be ported over to this model, with clear expectations of the resulting dynamics. An increase in the number of independent tissue expressions increases the rate of nonfunctionalization relative to subfunctionalization and therefore would be expected to increase the selective effect of the dosage barrier.

Another component that has been ignored to date is the deus ex machina process of neofunctionalization. Adding neofunctionalization, depending upon the associated assumptions about functional redundancy, has the potential to change the dynamics described here. The addition of neofunctionalization will create a fuller mechanistic model for duplicate gene fates, but is beyond the scope of the study here. There is complexity in identifying reasonable assumptions for the mutational process leading to neofunctionalization, which is why this is not as straight forward as it might appear.

While other levels of biological complexity in the underlying population genetics and molecular evolutionary processes are conceivable to model (including *trans*-effects), it is important to recognize the change in dynamics and process associated with just adding dosage balance to the characterization of the subfunctionalization process.

Conclusions

The complex dynamics of the interplay between subfunctionalization and dosage balance leads to opposite expectations for the timing and probabilities of retention for genes that encode proteins that function as multimeric complexes compared to those that function as monomers or homomultimers between whole-genome duplication and smaller-scale duplication. For proteins that function in multimeric complexes, retention following whole-genome duplication events through subfunctionalization is expected to be a non-neutral process.

Abbreviations

SSD	Small-scale duplication
WGD	Whole genome duplication
K_{eq}	Equilibrium constant
N_e	Effective population size
Sub + Dos	Subfunctionalization + dosage model
Sub-Only	Subfunctionalization only model from Stark et al. [59]

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12862-023-02116-y.

Additional file 1: Figure S1.

Acknowledgements

The authors wish to thank other members of the Liberles Research Group for helpful comments as well as anonymous reviewers. AEW would like to thank Francisco McGee for help conceptualizing protein-protein interactions. AEW would also like to thank Megan Lynn Wilson for her mentorship in programming.

Author contributions

AE and DAL conceived the study. AE wrote all code and generated all analyses under the supervision of DAL. AE and DAL wrote this manuscript. Both authors read and approved the final manuscript.

Funding

AEW was paid by money made available to DAL from Temple University. The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Availability of data and materials

The computer programs in the C++ programming language used for the analysis here are available on github through the following link: https://github.com/aewilson96/Wilson_Liberles_2022.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA.

Received: 31 August 2022 Accepted: 20 April 2023
Published online: 03 May 2023

References

1. Ohno S. Evolution by gene duplication. Berlin Heidelberg: Springer-Verlag; 1970.
2. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 2005;3: e314.
3. Hughes T, Liberles DA. Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. J Mol Evol. 2008;67:343–57.
4. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005;102:5454–9.
5. De Bodt S, Maere S, Van de Peer Y. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 2005;20:591–7.
6. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, et al. Widespread genome duplications throughout the history of flowering plants. Genome Res. 2006;16:738–49.
7. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 2008;18:1944–54.
8. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, et al. Polyploidy and angiosperm diversification. Am J Bot. 2009;96:336–48.
9. Barker MS, Vogel H, Schranz ME. Paleopolyploidy in the *Brassicales*: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other *Brassicales*. Genome Biol Evol. 2009;1:391–9.
10. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of gene duplication in plants. Plant Physiol. 2016;171:2294–316.
11. Pootakham W, Sonthirod C, Naktang C, Kongkachana W, Sangsrakru D, U-Thoomporn S, et al. A chromosome-scale reference genome assembly of yellow mangrove (*Bruguiera parviflora*) reveals a whole genome duplication event associated with the *Rhizophoraceae* lineage. Mol Ecol Resour. 2022. https://doi.org/10.1111/1755-0998.13587.
12. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290:1151–5.
13. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 2006;16:805–14.
14. Jin G, Ma P-F, Wu X, Gu L, Long M, Zhang C, et al. New genes interacted with recent whole-genome duplicates in the fast stem growth of bamboos. Mol Biol Evol. 2021;38:5752–68.
15. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010;11:97–108.
16. Reams AB, Roth JR. Mechanisms of gene duplication and amplification. Cold Spring Harb Perspect Biol. 2015;7: a016592.
17. Morgan LV. Polyploidy in *Drosophila melanogaster* with two attached X chromosomes. Genetics. 1925;10:148–78.
18. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725–32.
19. Marsit S, Hénault M, Charron G, Fijarczyk A, Landry CR. The neutral rate of whole-genome duplication varies among yeast species and their hybrids. Nat Commun. 2021;12:3126.
20. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol. 2007;308:58–73.
21. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 2007;8:R209.
22. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 2009;60:433–53.
23. Maere S, Van de Peer Y. Duplicate retention after small- and large-scale duplications. In: Dittmar K, Liberles D, editors. Evolution after gene duplication. 2011. p. 31–56.
24. Mottes F, Villa C, Osella M, Caselle M. The impact of whole genome duplications on the human gene regulatory networks. PLoS Comput Biol. 2021;17: e1009638.
25. Krogan NJ, Hughes TR. Signals and systems. Genome Biol. 2006;7:313.
26. Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A. 2012;109:14746–53.

27. Banerjee S, Feyertag F, Alvarez-Ponce D. Intrinsic protein disorder reduces small-scale gene duplicability. DNA Res. 2017;24:435–44.
28. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. 2003;424:194–7.
29. Veitia RA. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics. 2004;168:569–74.
30. Liang H, Plazonic KR, Chen J, Li W-H, Fernández A. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet. 2008;4: e11.
31. Veitia RA. Exploring the etiology of haploinsufficiency. BioEssays. 2002;24:175–84.
32. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell. 2007;19:395–402.
33. Teufel AI, Liu L, Liberles DA. Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. BMC Evol Biol. 2016;16:45.
34. Konrad A, Teufel AI, Grahnen JA, Liberles DA. Toward a general model for the evolutionary dynamics of gene duplicates. Genome Biol Evol. 2011;3:1197–209.
35. Li J-T, Hou G-Y, Kong X-F, Li C-Y, Zeng J-M, Li H-D, et al. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). Sci Rep. 2015;5:8199.
36. Roux J, Liu J, Robinson-Rechavi M. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. Mol Biol Evol. 2017;34:2773–91.
37. Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in Buckler Mustard. Plant Cell. 2016;28:17–27.
38. Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, et al. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 2021;22:103.
39. Hughes T, Liberles DA. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol. 2007;65:574–88.
40. Edger PP, Pires JC. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 2009;17:699–717.
41. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 2008;148:993–1003.
42. Liang H, Li W-H. Functional compensation by duplicated genes in mouse. Trends Genet. 2009;25:441–2.
43. Teufel AI, Johnson MM, Laurent JM, Kachroo AH, Marcotte EM, Wilke CO. The many nuanced evolutionary consequences of duplicated genes. Mol Biol Evol. 2019;36:304–14.
44. Birchler JA, Riddle NC, Auger DL, Veitia RA. Dosage balance in gene regulation: biological implications. Trends Genet. 2005;21:219–26.
45. Fernández A, Tzeng Y-H, Hsu S-B. Subfunctionalization reduces the fitness cost of gene duplication in humans by buffering dosage imbalances. BMC Genomics. 2011;12:604.
46. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73.
47. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256–60.
48. Reis-Cunha JL, Valdivia HO, Bartholomeu DC. Gene and chromosomal copy number variations as an adaptive mechanism towards a parasitic lifestyle in *Trypanosomatids*. Curr Genomics. 2018;19:87–97.
49. Ascencio D, Diss G, Gagnon-Arsenault I, Dubé AK, DeLuna A, Landry CR. Expression attenuation as a mechanism of robustness against gene duplication. Proc Natl Acad Sci U S A. 2021;118: e2014345118.
50. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 1994;256:119–24.
51. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151:1531–45.
52. Stoltzfus A. On the possibility of constructive neutral evolution. J Mol Evol. 1999;49:169–81.
53. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005;169:1157–64.
54. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 2005;5:28.
55. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.
56. Lynch M, Force AG. The origin of interspecific genomic incompatibility via gene duplication. Am Nat. 2000;156:590–605.
57. Davis JC, Petrov DA. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2004;2:E55.
58. Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis, Oryza Saccharomyces* and *Tetraodon*. Trends Genet. 2006;22:597–602.
59. Stark TL, Liberles DA, Holland BR, O'Reilly MM. Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization. BMC Evol Biol. 2017;17:38.
60. Birchler JA, Bhadra U, Bhadra MP, Auger DL. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol. 2001;234:275–88.
61. Rabinow L, Nguyen-Huynh AT, Birchler JA. A trans-acting regulatory gene that inversely affects the expression of the white, brown and scarlet loci in *Drosophila*. Genetics. 1991;129:463–80.
62. Yang H, Shi X, Chen C, Hou J, Ji T, Cheng J, et al. Predominantly inverse modulation of gene expression in genomically unbalanced disomic haploid maize. Plant Cell. 2021;33:901–16.
63. Shi D, Jouannet V, Agustí J, Kaul V, Levitsky V, Sanchez P, et al. Tissue-specific transcriptome profiling of the *Arabidopsis* inflorescence stem reveals local cellular signatures. Plant Cell. 2021;33:200–23.
64. Birchler JA, Veitia RA. One hundred years of gene balance: how stoichiometric issues affect gene expression, genome evolution, and quantitative traits. Cytogenet Genome Res. 2021;161:529–50.
65. Veitia RA. Gene dosage balance: deletions, duplications and dominance. Trends Genet. 2005;21:33–5.
66. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A. 2005;102:9541–6.
67. Birchler JA, Veitia RA. Protein-protein and protein-DNA dosage balance and differential paralog transcription factor retention in polyploids. Front Plant Sci. 2011;2:64.
68. Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA. Evaluating dosage compensation as a cause of duplicate gene retention in *Paramecium tetraurelia*. Genome Biol. 2007;8:213.
69. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006;444:171–8.
70. Kondrashov AS, Crow JF. A molecular approach to estimating the human deleterious mutation rate. Hum Mutat. 1993;2:229–34.
71. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019;12:315.
72. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;281:60–3.
73. Shubsda MF, McPike MP, Goodisman J, Dabrowiak JC. Monomer-dimer equilibrium constants of RNA in the dimer initiation site of human immunodeficiency virus type 1. Biochemistry. 1999;38:10147–57.
74. Green NM. Avidin. Adv Protein Chem. 1975;29:85–133.
75. Mall GK, Chew YC, Zempleni J. Biotin requirements are lower in human Jurkat lymphoid cells but homeostatic mechanisms are similar to those of HepG2 liver cells. J Nutr. 2010;140:1086–92.
76. Albe KR, Butler MH, Wright BE. Cellular concentrations of enzymes and their substrates. J Theor Biol. 1990;143:163–95.

Publisher's Note