universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Functional diffusion: A model for transcription factor binding site evolution and the origins of more complex regulatory structures

verfasst von | submitted by

Aleksa Spasić

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von | Supervisor:	Univ.-Prof. Dr. Joachim Hermisson

Acknowledgements

I would like to thank my supervisor Joachim Hermisson for mentoring me through this project, as well as Günter Wagner and Mihaela Pavličev for their helpful advice on the biological reality of the model.

I took Günter Wagner's advice on making enemies to hart in this and hopefully future work.

Statement of originality

This thesis incorporates segments of work previously undertaken during the Research Internship I course, under the supervision of Joachim Hermisson. Segments of the work in this thesis were also used in the Grant Proposal seminar course.

Contents

1 Abstract

Due to the finite nature of biological populations and the fact that the majority of mutations acting on a trait are deleterious, natural selection is limited in achieving the most optimal reachable design of a trait in many environments. Natural populations differ in the amount of genetic drift they experience (the long term effective population size [N_e] values vary by 5 orders of magnitude from higher Metazoans to Prokaryotes) and are subject to different mutation rates (Lynch et al. 2016). Based on these and other well established population genetics principles, an elaborate theory of genome evolution has been developed (Lynch, 2007; Koonin, 2009; Lässig, 2007) and is now further extended to explain more general cellular features (Lynch and Trickovic, 2020; Lynch, 2024). In this paper a model is developed of how transcription factor binding sites (TFBS) that control the expression patterns of genes evolve under two mutational processes (point mutation and duplications) under certain population genetic conditions. It is shown that, when selection is operating under a nearly effectively neutral regime ($N_e \cdot$ s is between 1 and 10), TFBS can enter a matching state from which a duplication event has a slight positive fitness effect. This leads to weak selection favouring a double binding site over a single one and increases the complexity of the regulatory structure without any new function initially arising. This process of creating a more complex regulatory region resulting from a suboptimal starting TFBS described here is called functional diffusion.

1.1 Zusammenfassung

Aufgrund der limitierten Größe natürlicher Populationen und der Tatsache, dass die Mehrheit der Mutationen, welche ein genetisches Merkmal betreffen, sich negativ auswirken, ist natürliche Selektion in den meisten Fällen eingeschränkt im Erreichen der optimalen Fitness für das betreffende Merkmal. Natürliche Populationen unterscheiden sich in der Menge an genetischen Drift welchen sie erfahren (die langfristige effektive Populationsgröße [Ne] variiert um fünf Größenordnungen zwischen höheren Metazoa und Prokaryoten), wie auch in ihren Mutationsraten (Lynch et al. 2016). Basierend darauf und weiteren etabliierten populationsgenetischen Prinzipien wurde eine umfassende Theorie der Genomevolution entwickelt (Lynch, 2007; Kroonin, 2009; Lässig, 2007) und kürzlich weiter ausgebaut um generelle zelluläre Eigenschaften erklären zu können (Lynch und Trickovic, 2020; Lynch, 2024). In dieser Arbeit wird ein Modell entwickelt das beschreibt wie sich Transkriptionsfaktor Bindungsstellen (TFBS), welche die Expression von Genen steuern weiterentwickeln unter dem Einfluss von zwei verschiedenen Mutationsmustern (Punktmutation und Duplikation) und unter bestimmten populationsgenetischen Bedingungen. Es wird gezeigt, dass wenn Selektion unter einem fast effektiv-neutralen Regime agiert ($Ne \cdot s$ ist zwischen 1 und 10), TFBS einen Zustand der Übereinstimmung mit der Transkriptionsmaschinerie erreichen, von dem aus eine Duplikation einen leicht vorteilhaften Effekt auf deren Fitness hat. Dies führt dazu, dass schwache Selektion doppelte Bindungsstellen, einfachen gegenüber bevorzugt und dadurch die Komplexität der regulatorischen Struktur erhöht, ohne das Aufkommen einer neuen Funktion dieser. Der Prozess der Entwicklung einer komplexeren regulatorischen Region aus einer suboptimalen TFBS wird hier als funktionale Diffusion bezeichnet.

2 Introduction

It is a well known result from theoretical population genetics that the ability of selection to purge deleterious mutations and promote advantageous mutations depends on the effective population size (N_e) (Kimura, 1962; Rice, 2004). In particular, in larger populations selection is more efficient in reaching and maintaining optimal design of a trait. This theoretical insight has been used to explain a myriad of phenomena in molecular evolution, such as amino-acid substitution rates in proteins (Ohta, 1973), enzyme activity (Hartl et. al, 1985), large scale differences in genome architecture (Lynch, 2007), mutation rates across the tree of life (Lynch et. al, 2016) and mRNA splicing accuracy (Bénitière et al. 2024). A particularly provocative hypothesis was proposed by Michael Lynch (2007a, 2007b) who states that many aspects of genomic complexity observed in higher Metazoa are the consequence of those species evolving under a much lower effective population size and thus allowing a proliferation of many more slightly deleterious fixation events. Later theoretical models were developed to incorporate more broader cellular phenomena in a similar fashion (Lynch, 2024). A fundamental aspect of this theoretical outlook is that many biologically interesting properties such as complexity, robustness, modularity and evolvability of higher Eukaryotes do not arise as a product of selection directly preferring such design, but rather are a byproduct of particular population-genetic conditions under which these populations evolve.

Transcription factor binding sites have been a topic of evolutionary research in both empirical (Nourmohammad and Lässig, 2011; Farley et al., 2015) and theoretical (Lässig, 2007; Tuğrul et al., 2015) domains. Biophysical models of transcription factor and binding motif interactions have been adapted to evolutionary genetics of binding sites to understand the dynamics of binding site evolution (Lynch and Hagner, 2015; Tuğrul et al., 2015; Lässig, 2007) under point mutations, deletion and insertion events. Given that the explosion of cellular and morphological complexity corresponds to the expansion of the non-coding part of the genome and a more complex regulatory structure (Carroll, 2008; Eddy, 2012), a deeper understanding of these elements from an evolutionary genetics perspective is warranted.

Here, a theoretical population genetic model is formed, of transcription factor binding site (TFBS) evolution under two mutational processes (point mutation and duplications). It is shown that under specific population genetic conditions (weak selection maintaining the TFBS sequence) binding sites can degenerate enough that a duplication event of the binding site can lead to a substantial fitness increase. The duplicated binding site state is realised only because selection initially was not able to maintain the optimal sequence in

the single binding site state.

2.1 Biophysics of TF binding

Transcription and expression are biophysical processes where a transcription factor (TF) binds to a particular transcription factor binding site (TFBS) and by doing so activates (increases expression) or suppresses (decreases expression) of a gene that it is regulating. Binding sites are short, usually around 10-16 nucleotides (Stewart et al., 2012) (closer to the lower end in eukaryotes and higher end in prokaryotes). TFBS are located in the cis-regulatory region of the gene (Carroll, 2008), which usually lies upstream of the gene and has a length of a few dozen to a few thousand nucleotides. The interaction between TF proteins and DNA can be position-unspecific attraction and specific attraction (Lässig, 2007). Position unspecific attraction is due the electrostatic interactions between the negatively charged DNA backbone and the positively charged TF protein molecule, while the specific attraction is due to specific properties of the locus (the sequence of the TFBS) and is achieved by hydrogen bonds between the protein and DNA. Theoretical and experimental studies (Mustonen et al., 2008) have elucidated some important properties on the nature of TF binding (Lässig, 2007):

1) The single nucleotides in the binding locus $A = (a_i...a_l)$ approximately contribute independently to the binding energy. $E(A) = \sum_{i=1}^{l} \epsilon_i(a_i)$

2) At each position i there is typically one nucleotide a_i^* where $\epsilon_i(a_i^*) = \min_a \epsilon_i(a)$. This means that for any TFBS sequence there exists one specific string of nucleotides $A^* = (a_1^*...a_l^*)$ for which the binding energy is lowest (i.e has strongest binding).

3) Mismatches with respect to the A^* have a energy cost of around 2 k_bT per nucleotide.

4) A two-state approximation can be made that for each site i in the TFBS sequence there exists only one nucleotide with the lowest energy (match) and the 3 other nucleotides have equal higher binding energies (mismatches).

Under these assumptions (Lässig, 2007; Lynch and Hagner, 2015) an equation can be derived that calculates the probability of a TF protein to be bound to its respective TFBS. That equation takes the form of (Lynch and Hagner, 2015):

$$P_{\text{on}}(m) = \frac{1}{1 + B \cdot e^{-b \cdot m}} \tag{1}$$

Here P_{on} is the probability that a TF is bound to the site, $b \cdot m$ is the mesure of the total strength of binding (m is the number of matches in a TFBS sequence with a range form 0 to l, b is the mismatch cost of 2 k_bT). The parameter B is described by the equation $B = \frac{G}{N_{\text{TF}} - N_{\text{off}}}$ (where G is the genome size, N_{TF} is the number of TF proteins in the cell and N_{off} is the number of TF proteins that bind to spurious off target sites) (Lynch and Hagner, 2015). In order to better understand the dynamics of TF binding and how it depends on the sequence of the TFBS with certain parameters, numerical evaluations were done in R to showcase some basic properties (Figure 1).

Notice that the expression $B \cdot e^{-b \cdot m}$ approaches zero as the number of matches (m) goes to infinity, and consequently $P_{\text{on}} \rightarrow 1$. More importantly, a key transition point in the binding property is obtained when $B = e^{b \cdot m}$, where the probability of binding is $P_{\text{on}} = 0.5$, marking the inflection point of the sigmoid curve (Figure 1). The position of this transition point is a function of the level of interference B. The higher the value B is, the more matching nucleotides are required to have a high binding probability. This can be easily interpreted when considering that B is larger for big genomes, since then the level of non-specific binding to DNA is increased. Another important thing to notice in this equation is that the probability of binding is not a direct function of the length of the binding site (l). However, the length of the site determines the highest possible matching state ($m_{\text{max}} = l$) since the TF protein only recognises the nucleotides within the specified region. For low values of B and high values of l it is possible that the binding probability approaches 1 and still have quite a few mismatches in the whole sequence (l) (Figure 1).

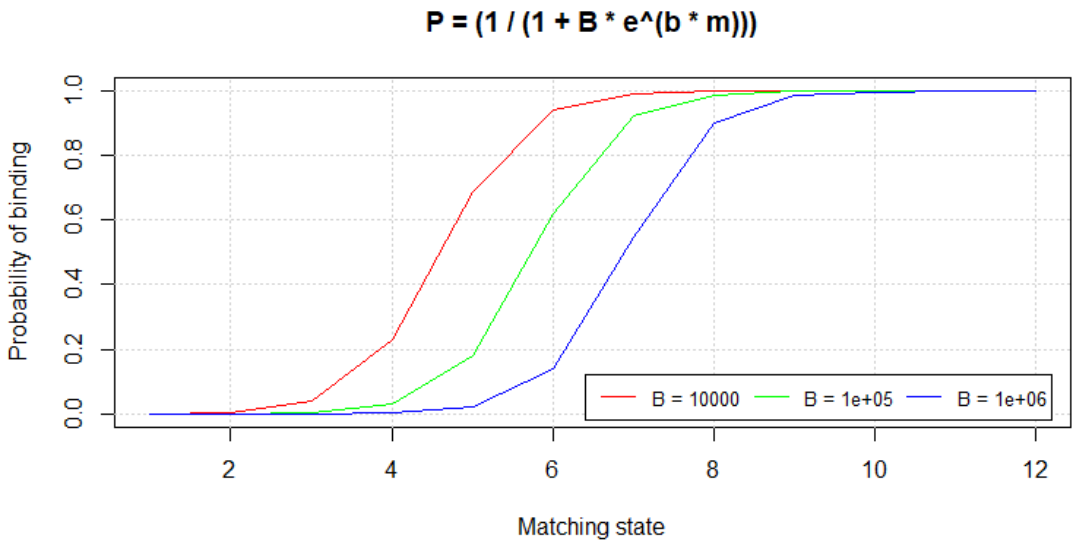

Figure 1: **Probability of binding as a function of the number of matches (matching state) in the TFBS sequence.** For low levels of B (red line), even for small matching states the binding probability can be close to 1. As the level of interference increases, so does the number of matches needed for high-probability binding.

2.2 Evolution of TFBS sequence

Based on this basic biophysical model of TF binding a population genetic model for the evolution of the binding sequence can be developed. In the model by Lynch and Hagner (2015) only point mutations, selection, and drift are considered. A haploid population size N (equal to the effective population size N_e) is assumed in the model (for a diploid population $2N$ needs to be substituted for N in the model). The mutation rate u is assumed to be equal between different nucleotides (A, T, C, G). Under this assumption, for each position in the TFBS sequence, it is 3 times more likely to observe a mismatch than a match in the sequence if selection does not maintain optimal binding (i.e., there is a mutational bias to more sub-optimized states). A fixed state approximation is also used in the model. This means that for the majority of time the population is in a monomorphic state (no variation in the TFBS sequence is present), and that the sequence can switch to only adjacent matching states (for example from matching state 7 to 6 or 8) in short transition (polymorphic) periods. This assumption is valid if and only if the waiting time between mutations that are destined for fixation (the inverse of the fixation rate $r = 2 \cdot N_e \cdot u \cdot s$) is much larger than the expected time for each new mutation to reach fixation ($\frac{ln(N \cdot s)}{s}$) (Desai and Fisher, 2007; Hermisson and Pennings, 2005). Under such assumptions, it can be shown that the distribution of matches in a TFBS of an arbitrary length l is (Lynch,

2024; Lynch and Hagner, 2015):

$$\Psi = C \cdot 3^{l-m} \cdot \binom{l}{m} \cdot e^{2 \cdot N_e \cdot s(m)} \tag{2}$$

Here C is a normalisation constant (equal to the reciprocal of the sum of all terms on the right of C for all values of m), N_e is the effective population size, W(m) is the fitness function relating the matching state of an individual to its relative fitness, m is the matching state (number of matches in the TFBS) and l is the length of the binding site.

The fitness function W(m), Lynch and Hagner (2015) assumed in their model is:

$$W(m) = 1 + s(m) = 1 + a \cdot P_{\text{on}}(m) \tag{3}$$

Here a is the total fitness advantage of having the gene fully expressed, P_{on} is the probability of binding given a specific matching state (equation 1). This simple model assumes that the maximal fitness is achieved when the gene is fully expressed (P_{on}=1)

Equation (2) gives the probability distribution of matching states for a TFBS under specific population genetic conditions (selection strength and population size). In order to get some initial understanding of the distribution for different population genetic conditions, I evaluated equation (2) for different values of the the product $N_e \cdot a$ (Figure 2). It can be seen how selection can optimise the sequence of the TFBS to be in highly matching states if the product of $N_e \cdot a > 1$. However, due to the mutation pressure in the opposite direction for low values of $N_e \cdot a$ mutation can override selection and the TFBS sequence degenerates (red line in Figure 2).

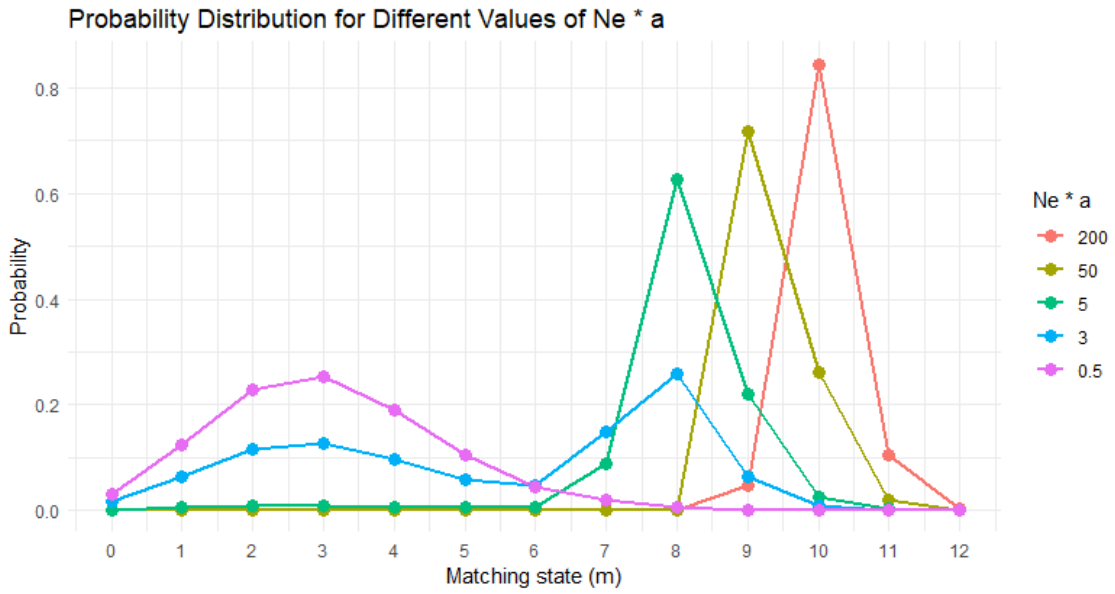

Figure 2: **The frequency of matching states for particular population genetic conditions.** The distribution reaches more and more suboptimal states as the product $N_e \cdot a$ decreases. For $N_e \cdot a < 1$ the distribution is the same as in the neutral case (effective neutrality).

If selection is extremely weak (the product $N_e \cdot a << 1$) then the value of $e^{2 \cdot N_e \cdot s_{(m)}} \rightarrow 1$, leaving just the left side of the equation (2). In this form the neutral distribution is obtained (Figure 3). This means that for populations with effective population sizes that are approximately smaller than $\frac{1}{a}$ no gene with a smaller value of the selection coefficient can be successfully maintained and expressed in the population.

What can be seen from this is that the value $e^{2 \cdot N_e \cdot s_{(m)}}$ is a scaling factor of the initial neutral distribution. Strong selection shifts the distribution of matching states to more highly optimised states. It has been shown that this scaling property is a general feature of linear fixed state models and the mathematical approach is vary similar to the one used in statistical physics (Sella and Hirsh, 2005; Lynch and Trickovic, 2020).

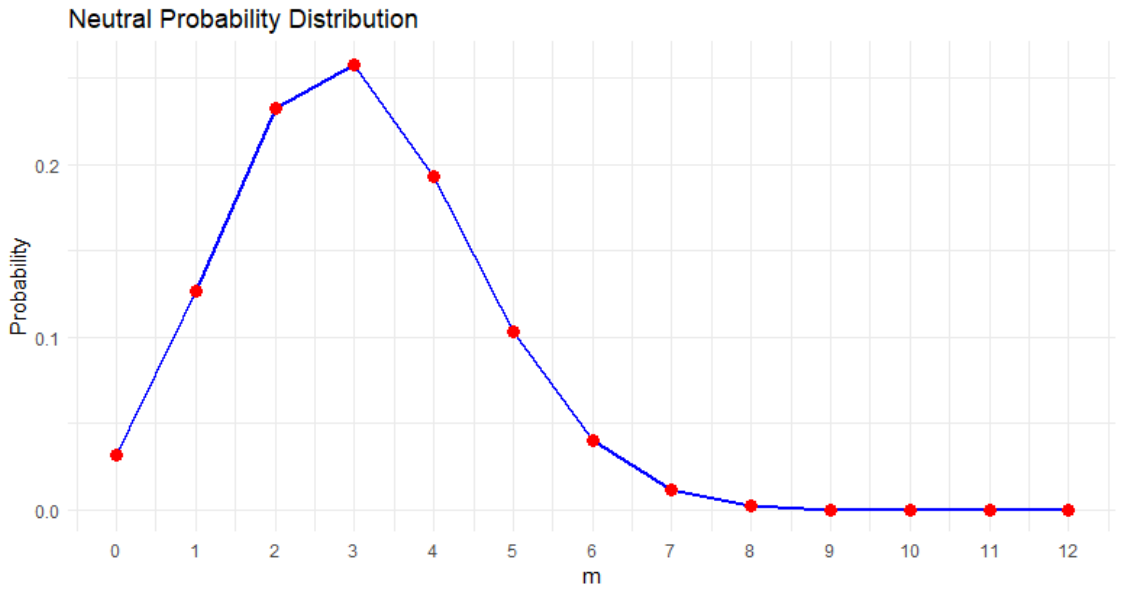

Figure 3: **The neutral frequency of matching states for $N_e \cdot a = 0$.** The matching state of the TFBS is far from the optimum (assuming $m_{\text{opt}} = l$). However, much more sequence change is possible then when compared to selection maintaining a highly optimised state.

3 Functional diffusion model

The functional diffusion model developed here illustrates how a more complex regulatory structure can emerge as a consequence of the system being maintained under weak selection and how sub-optimisation can lead to the emergence of adaptations that are not available in the optimised state.

Consider a TFBS evolving under a weak selection regime ($N_e \cdot a = 2 - 10$). Such a TFBS would be functional (not lost due to mutation pressure), but severely sub-optimised compared to the optimal matching state (assuming that $m_{\text{opt}} = l$). Given the sub-optimised state the TFBS is in, what would happen to the fitness of an individual that has a duplication of the binding site? In an optimised TFBS scenario the duplication event would most probably lead to a deleterious effect due to over-expression. However, if the expression level is far from the optimum in a sub-optimised TFBS scenario a duplication event could lead to a fitness increase by increasing the expression level to near optimum. In order to demonstrate this idea a model that incorporates the cost of over-expressing needs to be developed.

Following Tuğrul et al. (2015) it is assumed that the proxy for gene expression is the binding probability (P_{on}), or the sum of binding probabilities:

$$Expression_{\text{level}} \propto \sum d_i \cdot P_i \tag{4}$$

Where P_i is the binding probability (P_{on}) for TFBS i and d_i is a positional effect on expression. While P_{on} is solely determined by the matching state a binding site by equation (1), the parameter d_i can take any rational number value from 0 and upward, indicating how the position of the binding site relative to the gene and other binding sites can lead it to drive expression to a higher or lower level than the sequence suggests.

The simplest fitness function and initial model used here is a modification of a liner model originally used in the literature (Gerland and Hwa, 2002; Lynch and Hagner, 2015).

$$W = 1 - a \cdot |P_{\text{opt}} - P_{\text{real}}| \tag{5}$$

By using this modified equation where P_{opt} is the optimal value of the binding motif (assumed to be =1) and P_{real} is the realised value of the binding occupancy, overexpression is also penalised by leading to a selective disadvantage. P_{real} can be expressed as a sum of $d_i \cdot P_i$ where i is the number of TFBS regulating the expression level of a gene.

In the model it is also possible to consider that overexpression can have different fitness effect than underexpression by using different values of parameter a when $P_{\text{real}} \leq P_{\text{opt}}$ and when $P_{\text{real}} > P_{\text{opt}}$ (using a_1 for underexpression and a_2 for overexpression). Since underexpression leads to a partial loss of gene functionality and overexpression leads to energetic costs and ectopic expression (Lynch and Marinov, 2015; Farley et al, 2015) a partitioning of parameter a into the two categories is a reasonable assumption to model the underling biological reality.

To develop a model of the selective effect of TFBS duplication (a transition from a single binding site state to a two binding sites state) an explicit model from the general equation (5) is developed. The fitness value of a single TFBS influencing expression is

$$W_1 = 1 - a \cdot |P_{\text{opt}} - d_0 \cdot P_0| \tag{6}$$

Here P_{opt} is the optimal value of the binding motif and $d_0 \cdot P_0$ is the realised value (P_{real}) of the binding probability (given by eq. 1 for a specific matching state (m)). If $P_{\text{opt}} = d_0 \cdot P_0$, then $W_1 = 1$, corresponding

to the most fit possible genotype; if $d_0 \cdot P_0$ deviates from P_{opt}, then this leads to suboptimal values of fitness for those genotypes. It will always be assumed that $d_0 = 1$ so it can be dropped out of the equation. The positional influence of the initial TFBS is thus normalised to 1 and only the positional influence of the first duplicated binding site is considered (d_1).

The fitness value after a duplication event that created another identical binding site is:

$$W2 = 1 - a \cdot |P_{\text{opt}} - (P_0 + d_1 \cdot P_1)| \tag{7}$$

Where $(P_0 + d_1 \cdot P_1)$ is P_{real} after the duplication. Given that after the duplication the sequence of the two binding sites is the same ($P_0 = P_1$) we can simplify the the two fitness equations (for W_1 and W_2) by simply substituting P_0 with P_1 in both equations and by taking into account that d_0=1. The final form of the equations is then:

$$W_1 = 1 - a \cdot |P_{\text{opt}} - P_1| \tag{8}$$

$$W_2 = 1 - a \cdot |P_{\text{opt}} - (P_1 + d_1 \cdot P_1)| \tag{9}$$

The selective advantage/disadvantage of a TFBS duplication (s_{dup}) is $s_{\text{dup}} = W_2 - W_1$ (Lynch, 2012; Lynch and Hagner, 2015, Sella and Hirsh, 2005), substituting W_1 and W_2 with the equations above gives:

$$s_{\text{dup}} = a \cdot |P_{\text{opt}} - P_1| - a \cdot |P_{\text{opt}} - (P_1 + d_1 \cdot P_1)| \tag{10}$$

Given that a TFBS can duplicate from any of the starting matching states, using the equation for s_{dup}, I enumerated the fitness advantage/disadvantage of a TFBS duplication for a range of starting matching states in Figure 4.

As an alternative we can use a Gaussian fitness function (with $P_{\text{real}} = P_1$) in the starting case of one TFBS, after the duplication $P_{\text{real}} = P_1 + d_1 \cdot P_1$. Where P_1 is the realised binding probability for a single TFBS.

The Gaussian fitness function was also considered by Lynch and Hagner (2015) when modelling the matching state pattern with intermediate optimal expression. The function takes the form of:

$$W = \exp\left[-\left(\frac{P_{\text{real}} - P_{\text{opt}}}{\sigma}\right)^2\right]$$

The width of the Gaussian fitness function around the optimum is determined by the value σ. A higher σ value corresponds to a flatter and thus a more neutral fitness scenario. Under the Gaussian fitness model, the scaled selection parameter $\frac{N_e}{\sigma^2} = 1$ is the cutoff value that defines when selection stops total mutational breakdown (Lynch and Hagner, 2015)

A problem with studying the intermediate optimal expression level of with the Gaussian fitness function used by Lynch and Hagner (2015) is that the model assumes that over-expressing a gene has the same fitness effect as under-expressing it by the same magnitude. However, studies on gene expression and fitness have shown that often this is not the case (Keren et. al, 2016). If a gene does not reach optimal expression levels this has much grater fitness costs than if the expression level was to overshoot. In some cases it is also possible that over-expression has a larger fitness cost then under-expressing the gene. In order to incorporate these considerations into the model and have a more realistic picture of the relations between fitness and expression levels a modified Asymmetric Gaussian fitness function is used:

$$W = \left(e^{-\frac{(Preal-Popt)^2}{\sigma_1^2}} \cdot (Preal \leq Popt) + e^{-\frac{(Preal-Popt)^2}{\sigma_2^2}} \cdot (Preal > Popt) \right) \tag{11}$$

In this modified form it is possible to decouple the the fitness decline due to underexpression from the fitness decline due to overexpression. The left side of the sum is valid when P_{real} is smaller than P_{opt} ($(P_{\text{real}} \leq P_{\text{opt}}) = 1$) and the right side of the sum is valid when P_{real} is higher than P_{opt} ($(P_{\text{real}} > P_{\text{opt}})$=1). The value σ_1 corresponds to the fitness effects due to underexpression, and the values σ_2 to the fitness effects due to overexpression. By using two σ values for underexpression and for overexpression it is possible to consider different fitness regimes depending on which side of the optimum the organism finds itself.

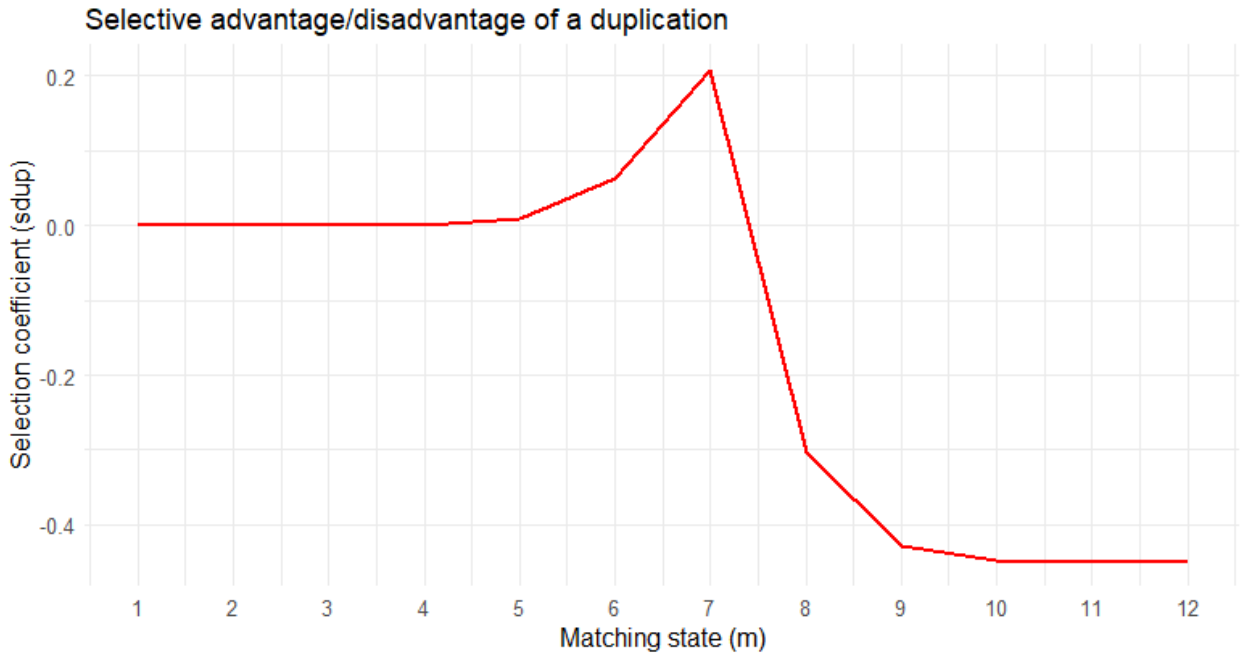

Figure 4: **The selective advantage/disadvantage of a TFBS duplication depending on the initial matching state that duplicated.** The image shows that there is selective advantage of the duplication event for intermediate values of matching, however if the site is already selectively optimized (a duplication from a high matching state) then this leads to a selective disadvantage due to over-expression (equation 7). Parameters: a=0.5, l=12, d1=0.9, Popt=1, B=10^6

.

3.1 Population genetic conditions for Functional diffusion

Given the fact that Ψ gives the distribution of times a TFBS is in a particular state and that the selective advantage/disadvantage of a TFBS duplication event can be calculated, it is possible to consider the population genetic conditions under which a duplication event will be favourable and a transition to a two binding site state is selectively achievable. In order to show this, numerical derivations under different population genetic conditions (different values of the $N_e \cdot$ s) were done in R. Figures 5 and 6 display the results, using the liner fitness model (equation 5) developed here. Under this fitness regime where under-expression and over-expression are equally penalised by selection the conditions for functional diffusion are meet only in the parameter range of $N_e \cdot$ s=1-10. For higher values of $N_e \cdot$ s, the distribution of TFBS matching states never overlaps with a matching state from which a duplication will be favourable, it mostly wanders in higher matching states from where a duplication event is always deleterious due to over-expression. This shows that functional diffusion is possible when certain population genetic conditions are meet and it seems that this is more often the case for species with a smaller effective population sizes such as Metazoans and is less likely

for species with high effective population sizes such as most bacterial species. In figures 5 and 6 only the linear fitness function is presented, for the Gaussian case (equation 11) see figures 8 and 9.

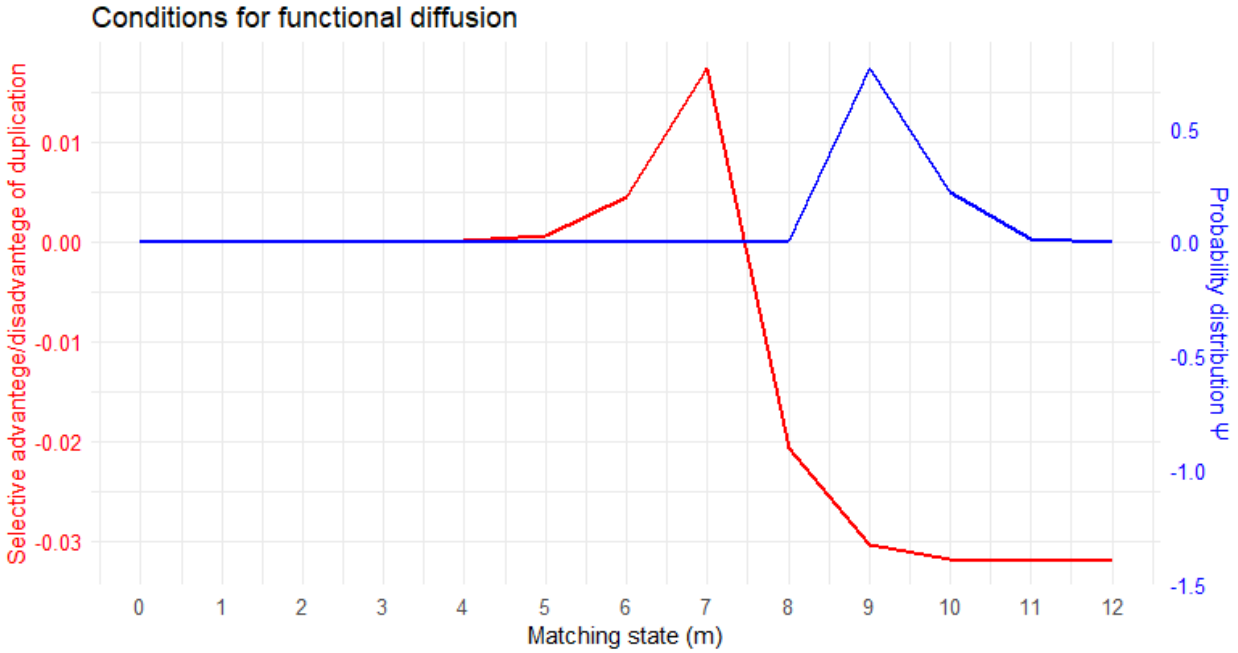

Figure 5: **The selective advantage/disadvantage of a TFBS duplication depending on the initial matching state that duplicated and the probability of being in the matching state.** Here the product of $Ne \cdot a = 40$ ($d_1 = 0.8$, $B = 10^6$, $l = 12$, $b = 2$). This leads to efficient selection acting on the sequence such that the distribution Ψ never enters a matching state from where a duplication is advantageous

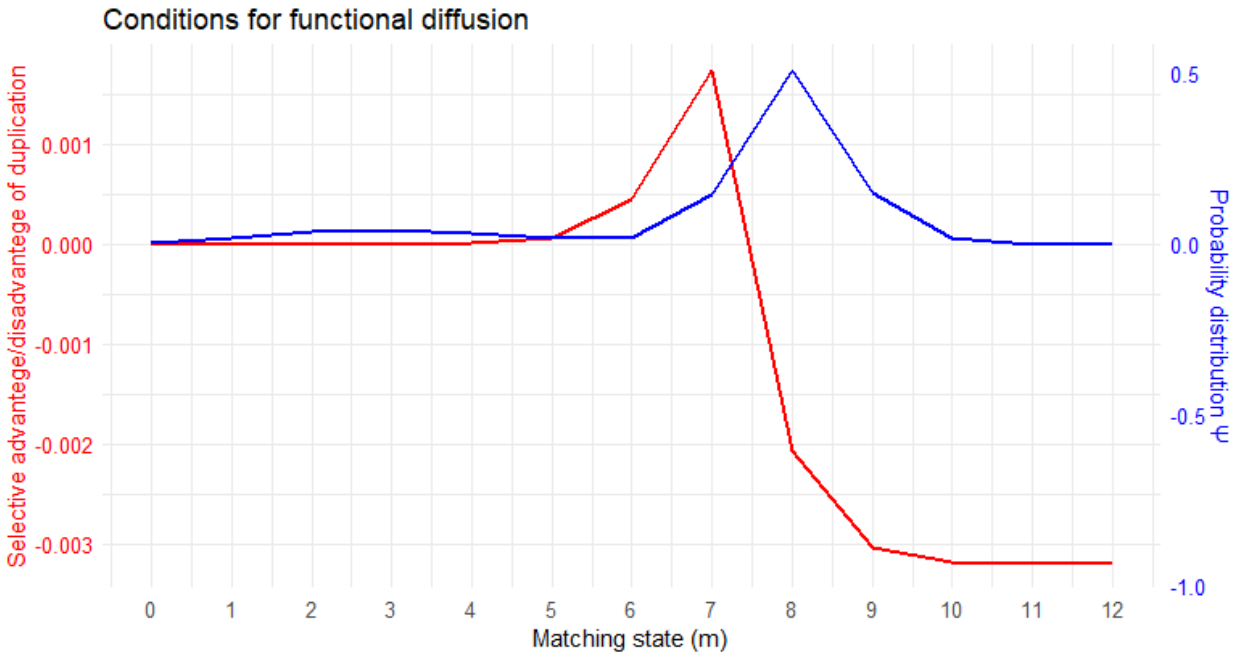

Figure 6: **The selective advantage/disadvantage of a TFBS duplication depending on the initial matching state that duplicated and the probability of being in the matching state.** Here the product of $Ne \cdot a = 4$ ($d_1 = 0.8$, $B = 10^6$, $l = 12$, $b = 2$). This leads to inefficient selection acting on the sequence such that the distribution Ψ enters a matching state from where a duplication is advantageous

This shows that even in very restrictive population genetic conditions when the cost of overexpression is equal to the cost of underexpression, a duplication of a TFBS from a specific suboptimised state is possible to be selectively favoured.

In order to test the theory developed here simulations were done in Phyton of the duplication process (Figure 7). A Wright-Fisher simulation was preformed using 5000 replicates of a population of size N=1000. The starting matching state of the population was sampled from the distribution given by equation (2). Duplications of the TFBS were introduced at a rate u_d. The populations evolved for 10000 generations during which fixation events were counted. The expected number of fixations ($Counts_{\text{fix}}$) from the theory developed here is simply expected to be proportional to the product of the fixation probability of the duplication

$$p_{\text{fix}} = \frac{1 - e^{-2 \cdot N_e \cdot s \cdot p}}{1 - e^{-2 \cdot N_e \cdot s}}, \tag{12}$$

(Kimura, 1962) from a specific matching state m_i and the probability of the population being in that specific matching state (Ψ_i):

$$Counts_{\text{fix}} \propto p_{\text{fix}} \cdot \Psi \tag{13}$$

Equation (13) has been also compared to the simulation results (Figure 7) and gives a good match. Most fixation events that happened are from the suboptimised state (m=7 in this case) from which a duplication event has the most positive fitness effect.

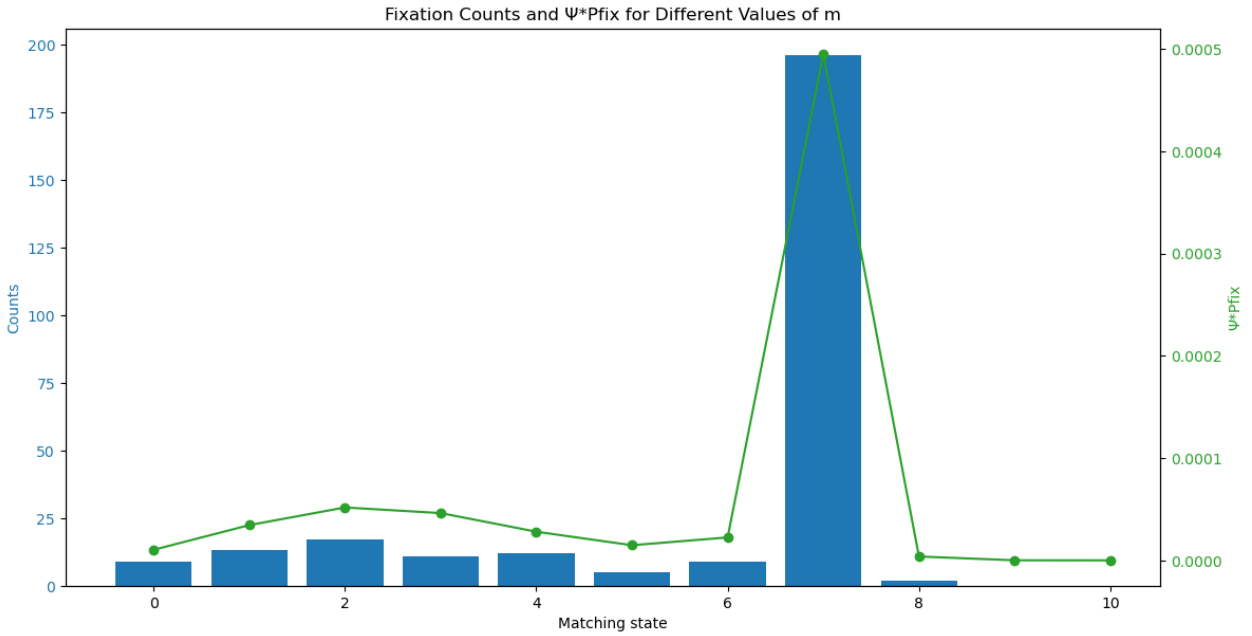

Figure 7: **Wright-Fisher simulations and the $Counts_{\text{fix}}$ prediction.** The simulations were done using the linear fitness function equation (4). With parameter values a= 0.005, N=1000, u_d = 0.00001, l=10, d1=1. The graph shows that the majority of fixation events happened from the suboptimized matching state m=7, which is in line with the theoretical prediction given by the green line (equation 13)

In order to study the effect of different selection coefficients penalising underexpression and overexpression the Gaussian model (equation 11) was used (Figures 8 and 9). In a similar parameter regime functional diffusion was found to be possible $\frac{N}{\sigma^2} = 2 - 10$. The Gaussian fitness function leads to similar effects to the linear model.

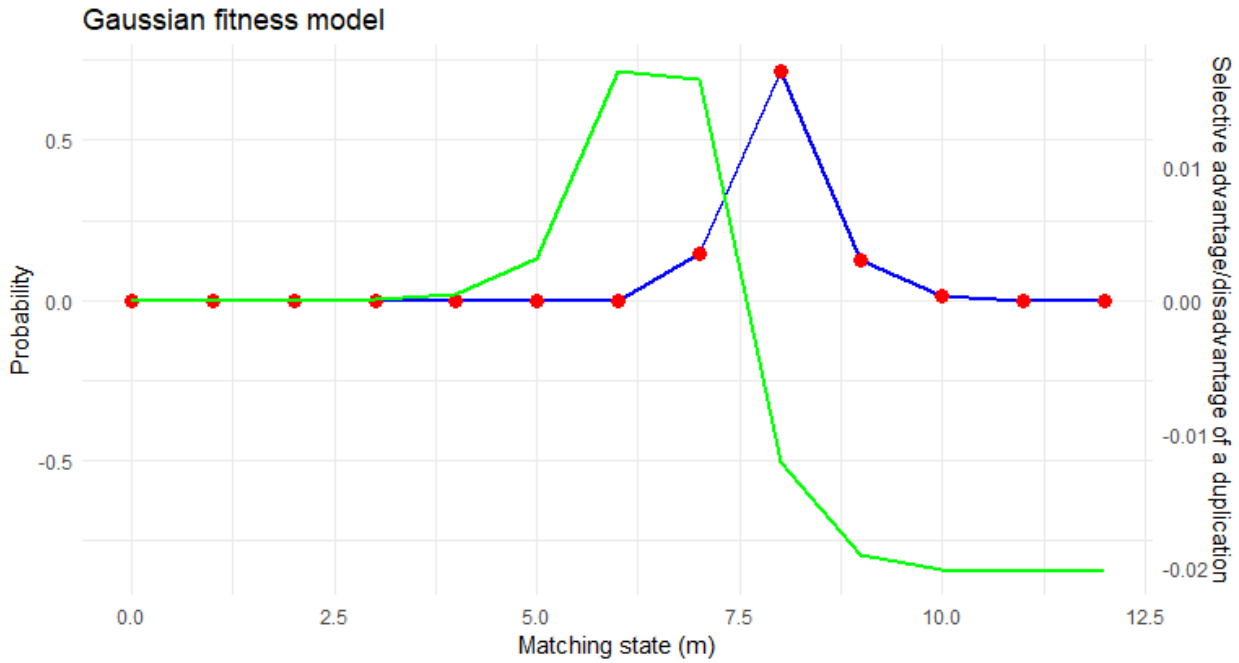

Figure 8: **Selective effect of a duplication under Gaussian fitness function.** Parameters: N=100, $\sigma_1 = 3.5$, $\sigma_2 = 7$, $B = 10^6$, $d_1 = 1$

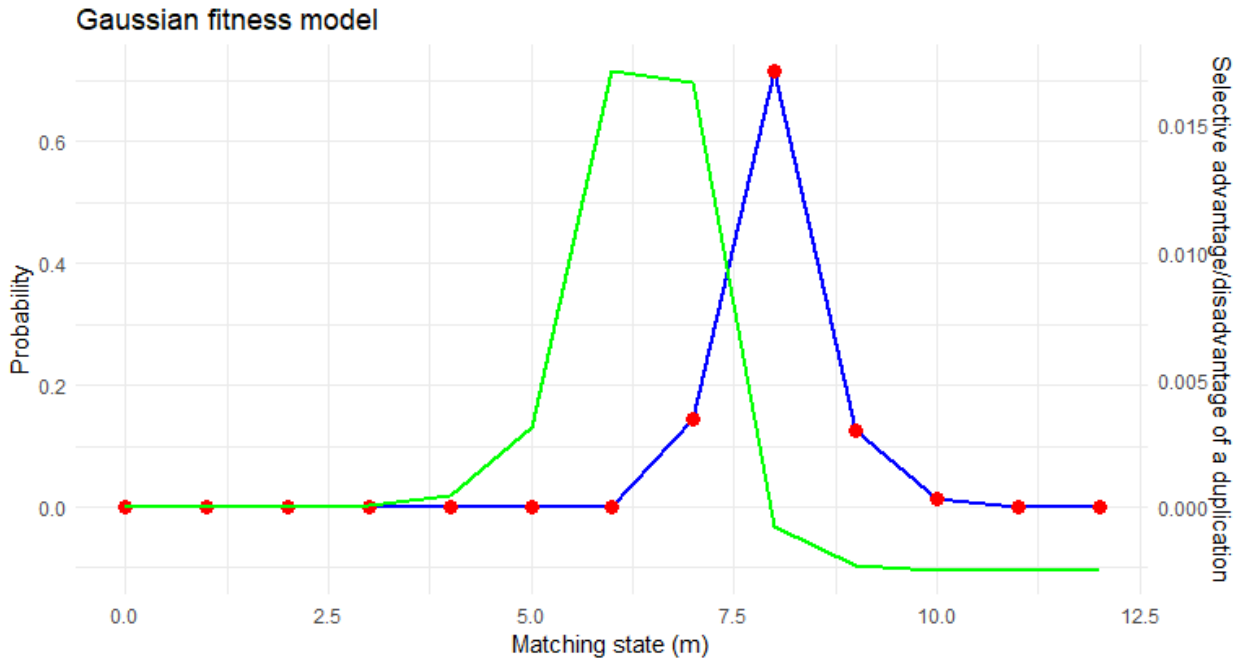

Figure 9: **Selective effect of a duplication under Gaussian fitness function.** Parameters: N=100, $\sigma_1 = 3.5$, $\sigma_2 = 20$, $B = 10^6$, $d_1 = 1$

In order to see that selection is effective in promoting the duplication event, the probability of fixation

equation (p_{fix}) (Kimura, 1962) was graphed with the matching state probability (Ψ). It shows that there is a substantial increase in the probability of fixation in the functional diffusion regime (Figure 10).

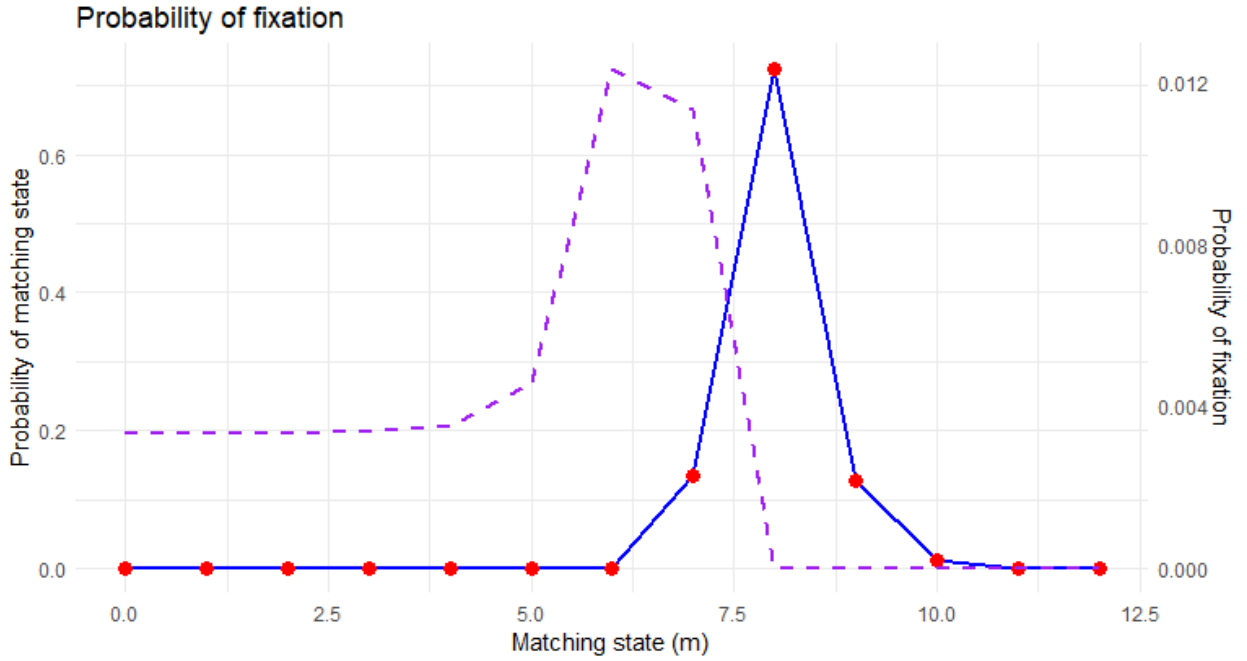

Figure 10: **Probability of fixation given a duplication event from a specific matching state and Ψ.** A Gaussian fitness function was used to calculate s_{dup} used in the Kimura equation for p_{fix}. Of particular importance is the shift in the probability of fixation when shifting from the most common type of matching state (m=8) to the suboptimized state (m=7). Parameters: N=300, $\sigma_1 = \sigma_2 = 6$

Of particular importance is the shift in the probability of fixation shown in Figure 10. This shows that when in a weak selection regime the matching state from which a duplication event is favourable is only one mutation away from the most common matching state observed in the population.

3.2 Waiting time for functional diffusion

In order to know the time scale on which this process is happening a waiting time equation for the origin of the duplicate has been derived. Taking into account that the rate of duplication for each matching class is $r_i = N_e \cdot p\text{fix} \cdot u_d$ the waiting time for a fixation event assuming a constant matching class is $\frac{1}{r}$ (Walsh and Lynch, 2018). Taking into account that the proportional amount of evolutionary time spent in each matching class is given by the distribution Ψ (equation 2) the final probability of fixation equation is given by:

$$t_{\text{dup}} = \frac{1}{N_{\text{e}} \cdot \text{u}_d \cdot \int_0^l p_{\text{fix}}(m)\Psi(m)dm} \tag{14}$$

In equation (14) every fixation probability of a duplication for each matching class is multiplied by the respective proportion of time spent in that matching class. The other parameters are the effective population size (N_e) and the rate of duplication (u_d).

Using the Gaussian fitness function with asymmetric effects on underexpression and overexpression I graphed this equation in order to display some important properties of the model (Figure 11 and 12). As shown in Figure 11, there is a decrease in the waiting time prior to reaching neutral rates of fixation as the binding sequence spends more time in the sub-optimised state. Interestingly, the fixation times after increasing as selection becomes more efficient at some point again start to fall of (even though the novel minimum never reaches the level of the original in the functional diffusion zone).

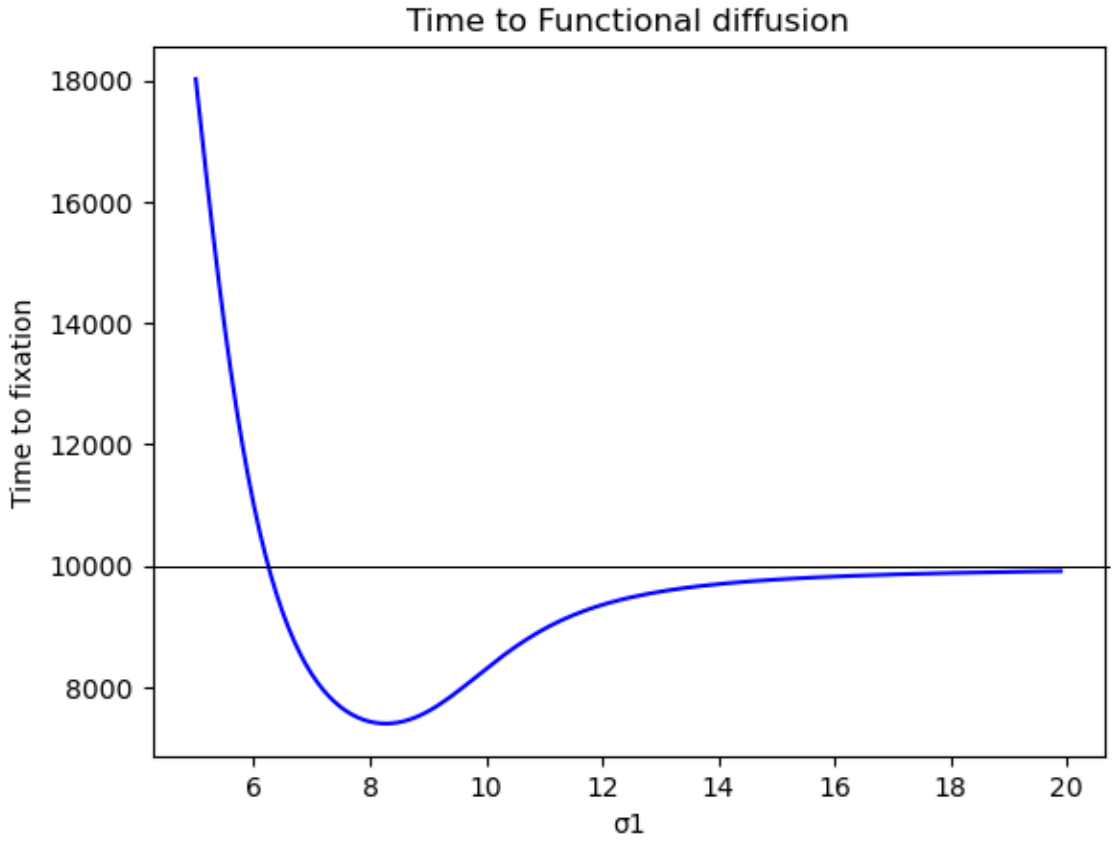

Figure 11: **The waiting time (t) is plotted on the y-axis and the selective cost σ_1 is plotted on the x-axis** (σ_1 increases correspond to the more neutral scenario). In the range of σ_1 from 6 to 12 a dip in the waiting time compared to the neutral expectation is observed. This region corresponds to the functional diffusion zone where the Ψ distribution enters a state form where a duplication is favoured. The neutral waiting time expectation is denoted by the black line at $\frac{1}{u_d}$ = 10000. Parameters Ne = 340, σ_2 = 15, B = 1000000, b = 2, u_d= 0.0001, l= 10

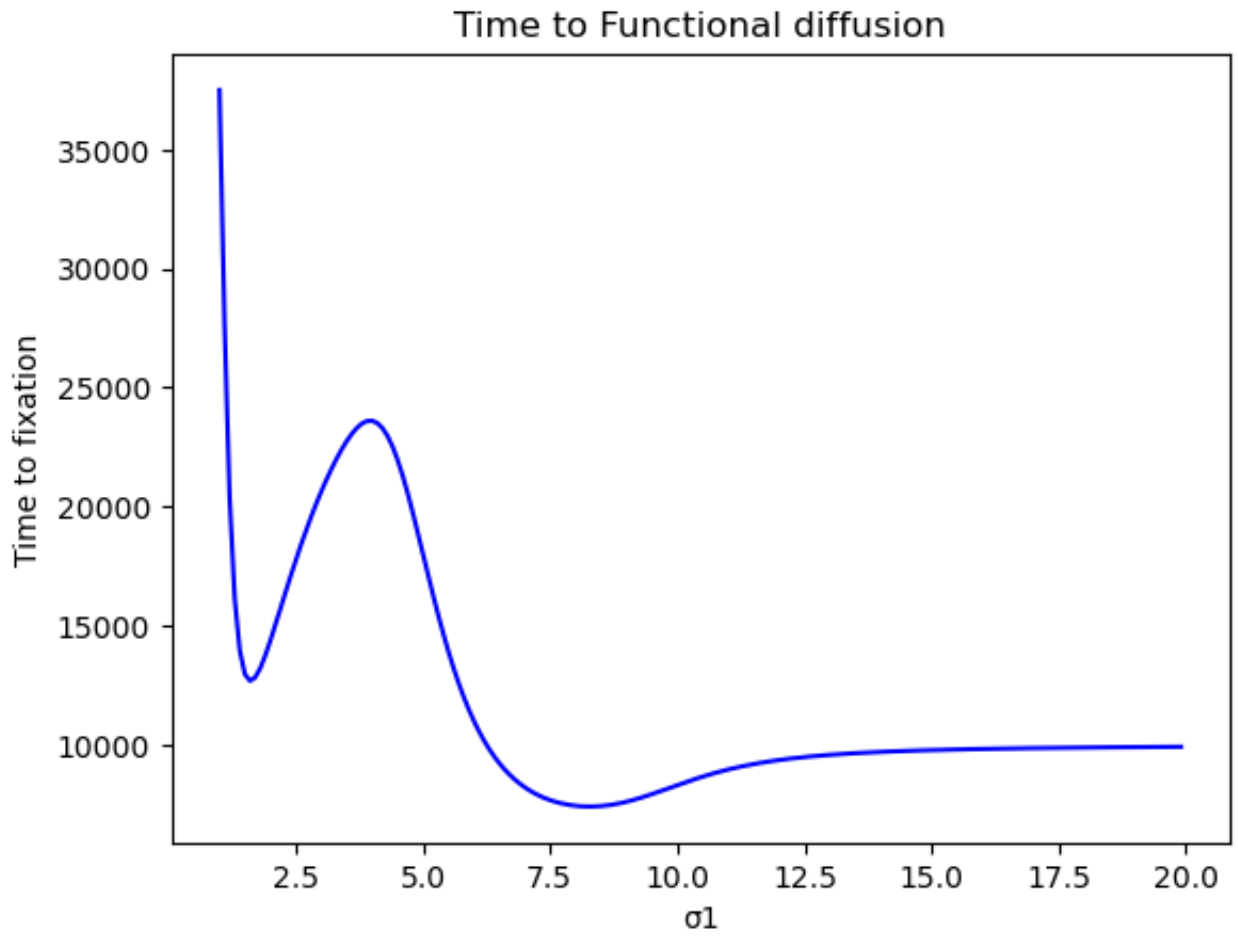

Figure 12: **A larger parameter zone for σ_1 for the waiting time (t)**. It shows a zone where increasing selection acting on a binding site decreases the waiting time again (u_d to selection being more efficient in promoting the duplication event) Parameters Ne = 340, c2 =σ_2= 15, B = 1000000, b = 2, u_d= 0.0001, l = 10

In order to gain a better understanding of the biophysical systems conditions necessary for functional diffusion I graphed the waiting time equation for different values of the binding site length (Figure 13). As the average length of the binding site gets lower, so does the waiting time for functional diffusion decrease. This is due tot he fact that for shorter binding sites in order to achieve high binding probability, selection needs to optimise almost every nucleotide of the binding sequence, however, for larger binding sites, a few mismatches are tolerated since high binding probability can be achieved while still having a few mismatches in the sequence (Figure 1).

An interesting observation (Figure 12) in this prediction is the initial increase and following decreases of the waiting time. A possible explanation is that, as selection grows stronger, at one point it stops changing the distribution Ψ but still increases the probability of fixation of a duplication (p_{fix}). This might explain the

decrease in the waiting time. However, once selection is too strong, it overcomes the mutational pressure and optimises the binding site at higher matching states, leading to the sharp increase in waiting time on the far left end of the graph.

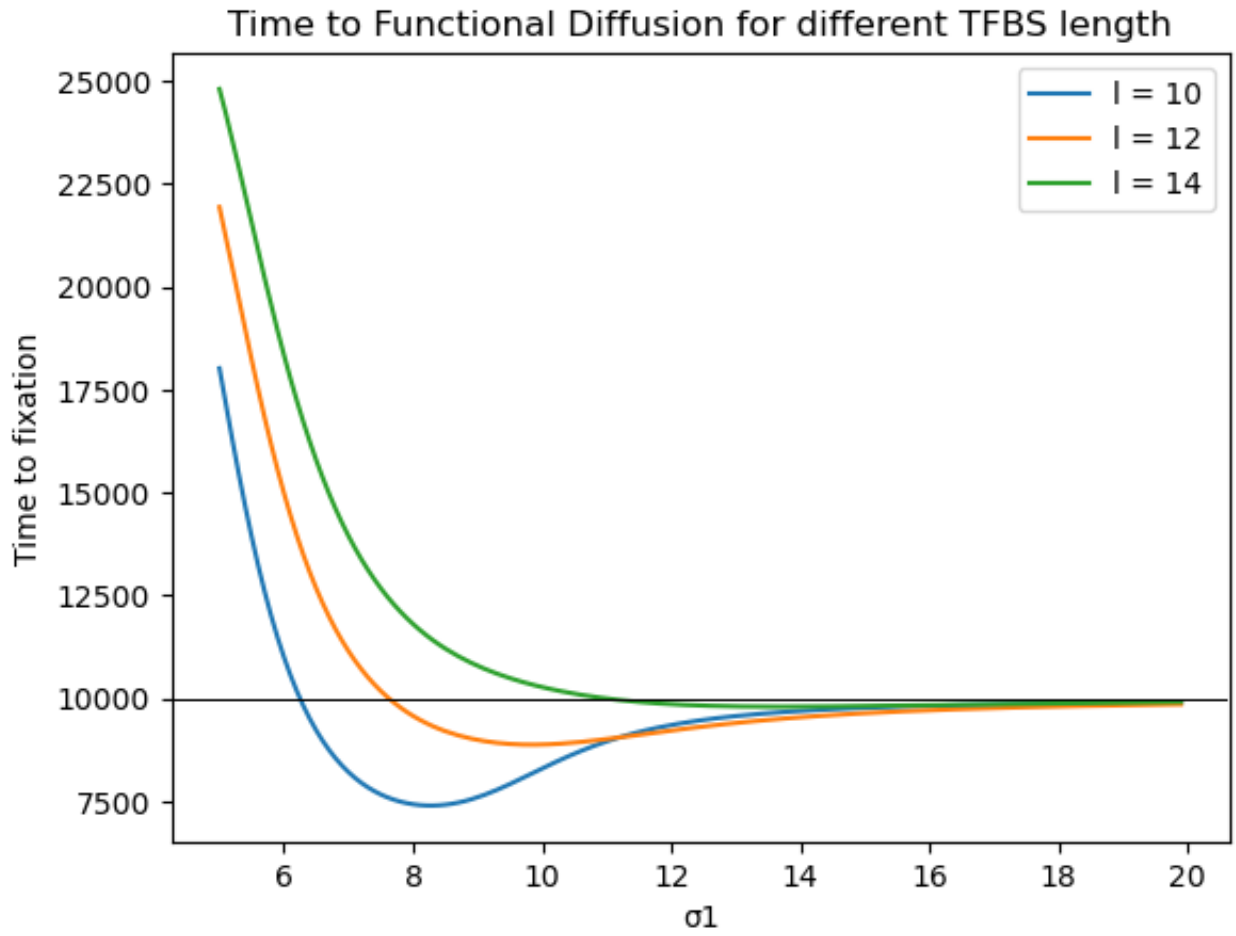

Figure 13: **The waiting time to the origin of a successfully duplication**. Shorter biding sites have lower waiting times in the functional diffusion zone. Parameters N_e = 340, σ_2= 15, B = 1000000, b = 2, u_d= 0.0001

3.3 Functional diffusion as a gene duplication model

While the initial model developed here is focused on TFBS duplications that control the expression level of a single gene, this model can also be reinterpreted as a model of whole gene duplications under complete linkage (no recombination). The model lies between previously developed models of neofunctionalization and subfunctionalizastion (Lynch, 2007), in that there is selection promoting the gene duplication event (similar to neofunctionalisation), but with no actual new function arising (similar to subfunctionalization). Previous empirical studies have shown a complicated relationship between subfunctionalisation and neofuctionalisation

(Sandve, 2018; He and Zhang, 2005), and such empirical investigations take into account expression levels for demonstrating sub- or neo- functionalisation events. It is possible that the basic idea behind the model developed here can shed light on some of these complications. However, to consider this, it is necessary to develop a model of gene duplication under functional diffusion that incorporates recombination events between the duplicates.

4 The Functional diffusion zone

Using the theory developed it is possible to determine the parameter zone under which functional diffusion is possible. By using the time to fixation equation (14) we can define the functional diffusion zone as the parameter space under which the average fixation time (t_{dup}) is lower than the the neutral expectation. In this parameter space the TFBS sequence is often suboptimsed enough that selection is efficiently promoting the duplication event.

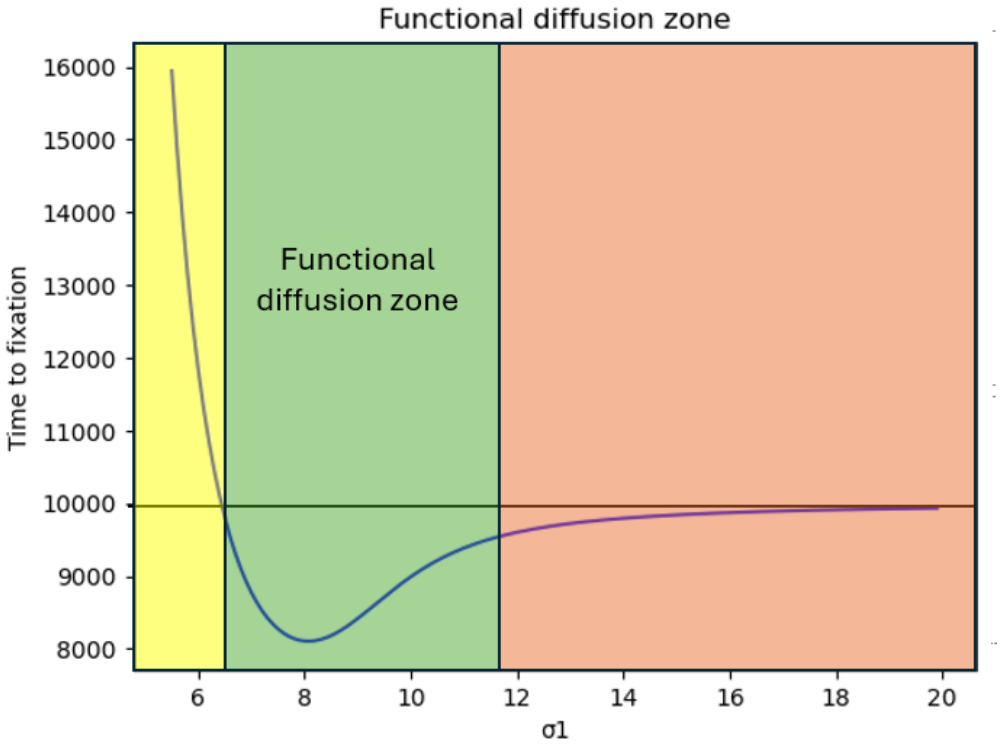

Figure 14: **The functional diffusion zone**. The green zone on the graph corresponds to the parameter zone where functional diffusion is a relevant process. In the orange zone the TFBS is totally lost by drift due to effective neutrality, and in the yellow zone the the binding site sequence is maintained at a too high of a matching state for functional diffusion to be a probable process. Parameters: $l = 10$, $N_e = 300$, $\sigma_2 = 10$, $d_1 = 1$, $B = 10^6$

Under a Gaussian fitness model (equation 11) with a not too high penalty for overexpression ($\frac{N_e}{\sigma_2} = 3$) and biophysical properties that are representative of higher Eukaryotes (lower lengths of TFBS (l) and higher values of the interference parameter (B)) (Lynch, 2024, Lynch and Hagner, 2015) we can see that there exists a range of parameter values where functional diffusion is possible. Because in population genetics the typical expression for for the strength of selection is usually represented by $N_e \cdot s$ in order get a similar intuitive scaling in the Gaussian model the equivocation can be made that $s_{\text{Gauss}} = \frac{1}{\sigma}$, then the parameter range for functional diffusion under the biological conditiones noted above is $N_e \cdot s_{\text{Gauss}} = 2.5 - 7.5$ (Figure 14). For values of $N_e \cdot s_{\text{Gauss}} > 7.5$ the selection optimises the binding sequence too much, and for values $N_e \cdot s_{\text{Gauss}} < 2.5$ the initial TFBS is lost due to drift.

4.1 Lower values of d1 increase the functional diffusion zone

The functional diffusion zone shown in Figure 14 takes into account only $d_1 = 1$. This would mean that all duplications create a binding site of equal effects on expression (and fitness). Numerical evaluations of the model show that if the duplicated bindings site had such an effect that it had a lower effect on fitness than the initial binding site, then the zone for functional diffusion is increased (Figure 15). This happens due to the fact that duplicated binding sites with $d_1 < 1$ can optimise the expression level better, with less overexpression effects than if $d_1 = 1$. The negative effects on fitness coming from a duplication from a more optimised state are also minimised in such a scenario, increasing the chances of fixation for such duplicates as well.

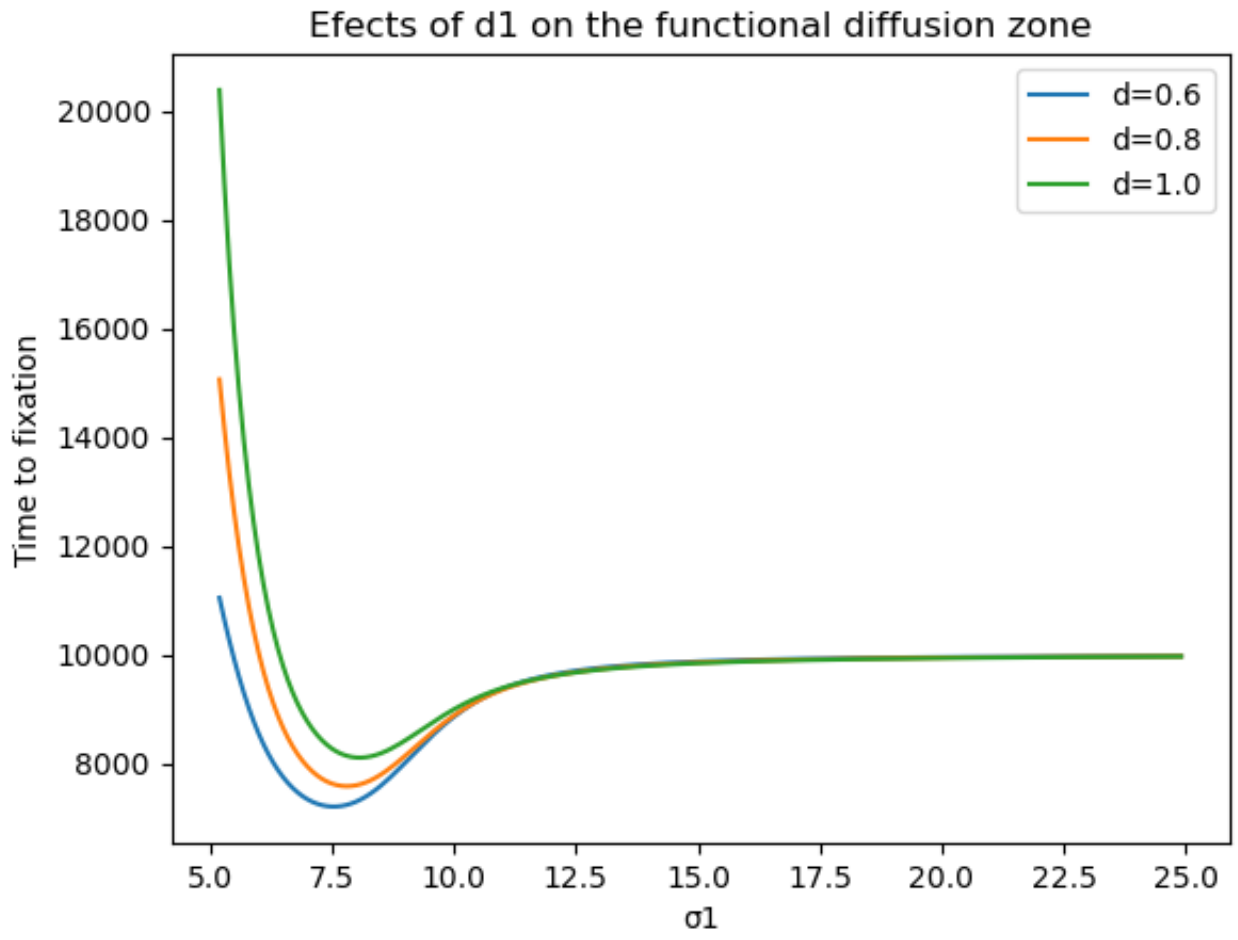

Figure 15: **The effects of parameter d1 on the functional diffusion zone**. Lower values of d_1 expand the functional diffusion zone. Parameters: $l = 10$, $N_e = 300$, $\sigma_2 = 10$, $B = 10^6$

This leads to an increase of the functional diffusion zone on the higher end of the $N_e \cdot s_{\text{Gauss}}$ product. Such that the new cut of for the functional diffusion zone can go up to and partially even exceed $N_e \cdot s_{\text{Gauss}} > 10$. Values of $d_1 > 1$ lead to a shrinking of the diffusion zone. High values of d_1 lead to too high of a penalty for overexpression that can overcome the fitness benefit of a the gene being functionally expressed.

4.2 Effects of parameter B on the functional diffusion zone

Parameter B introduced by Lynch and Hagner (2015) into their model is an intuitive representation of the levels of interference in transcription factor binding. The range of values the parameter B can take can be as low as 10^3 and as high as 10^7 (Lynch and Hagner 2015; Lynch, 2024). For low values of B the number of matches need for optimal binding is low and it increases as B increases (Figure 1). The effects of parameter

B are on the functional diffusion zone are shown in figure 16.

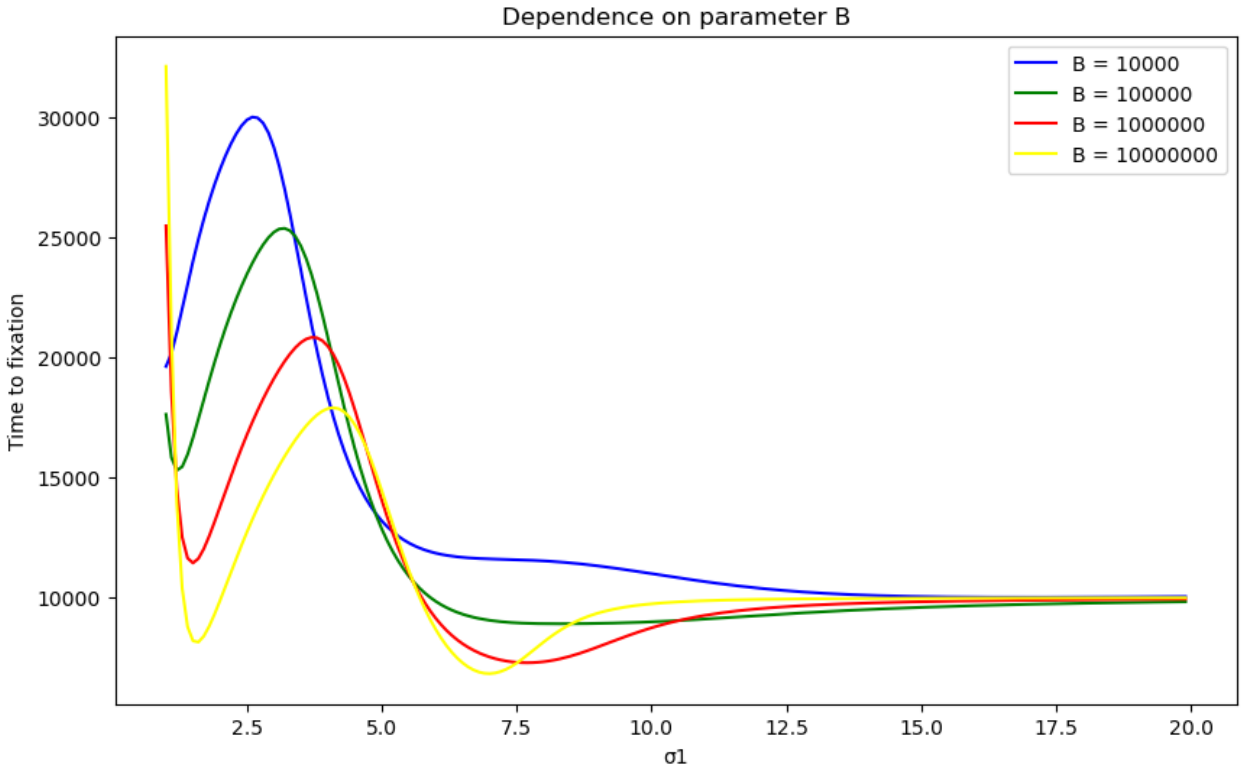

Figure 16: **The effects of parameter B on the functional diffusion zone**. Lower values of B decrease the functional diffusion zone. Parameters: $l = 10$, $N_e = 300$, $\sigma_2 = 10$, $d_1 = 1$

The functional diffusion zone totally vanishes for low values of the interference parameter B (Figure 16). This is because for low values of B, the sequence is already optimised at lower matching states (Figure 1) and no suboptimised state is ever present enough in the population for the duplication event to be favourable.

4.3 Effects of parameter b on the functional diffusion zone

There is considerable variation in the average cost of a mismatch mutation, but typically range between 1 and 3 k_bT (Lynch and Hagner, 2015; Lässig, 2007). Numerical evaluations of the time to fixation equation for different parameter values of b show that lower values of b favour functional diffusion, while higher values disfavour the process (Figure 17).

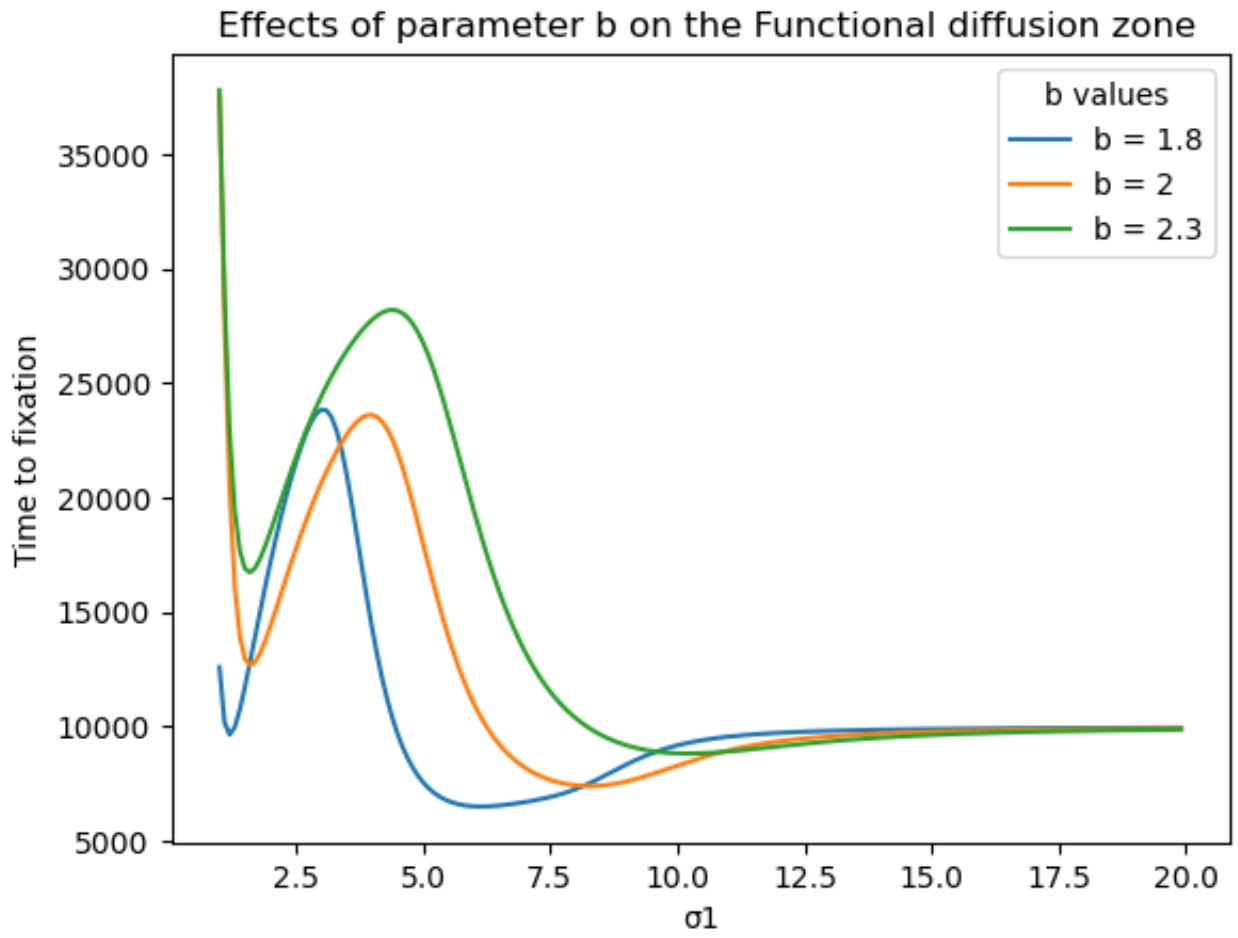

Figure 17: **The effects of parameter b on the functional diffusion zone**. Lower values of b increase the functional diffusion zone. Parameters: $l = 10$, $N_e = 340$, $\sigma_2 = 15$, $d_1 = 1$, $B = 10^6$

The reason this is so is because parameter b determines the gradient of change in the probability of binding equation (1). Lower values lead to a smoother line of change (smaller fitness differences between matching states) (Figure 18). Because of the smaller change in fitness between the matching states selection is further inhibited by mutation pressure to a suboptimised state.

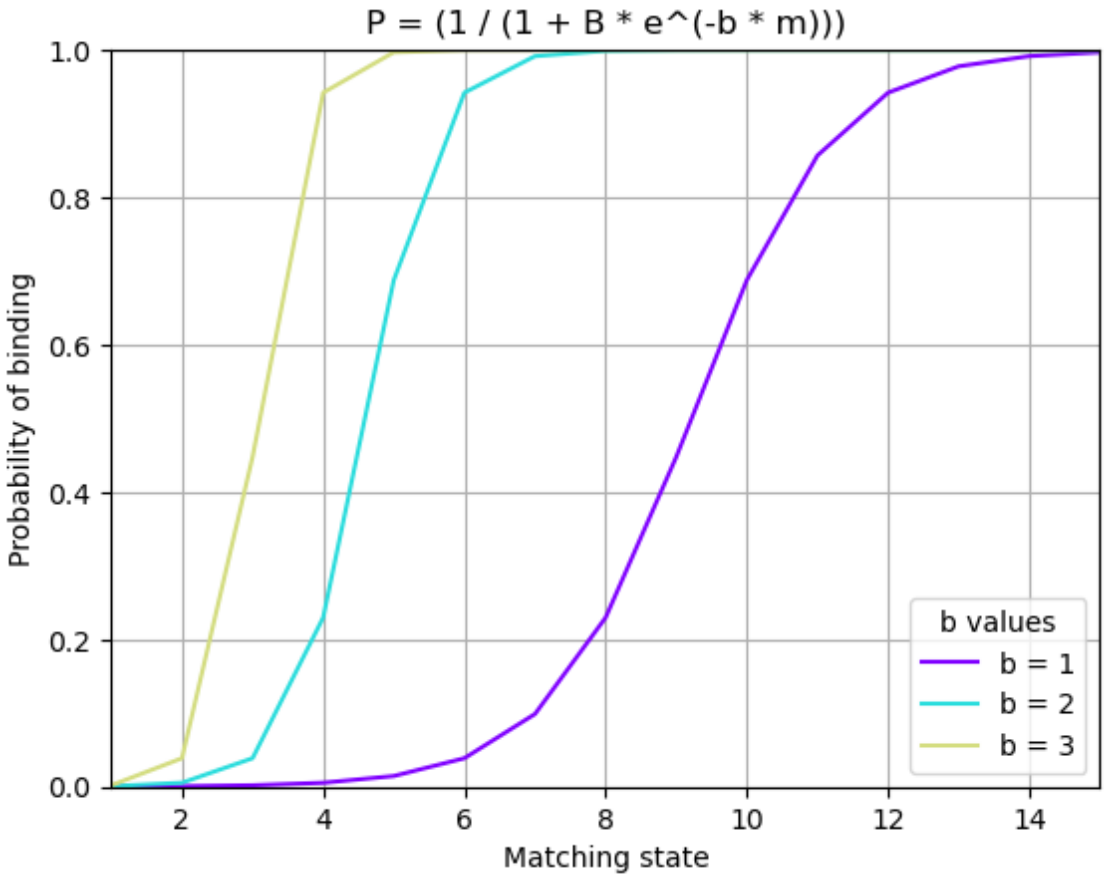

Figure 18: **The effects of parameter b on probability of binding (equation 1)**. Lower values of b lead to a smoother curve. Parameters: $B = 10^4$

5 Discussion and Implications

5.1 The systems conditions for functional diffusion

Whether and how often functional diffusion happens in nature and in which groups of organisms is it more probable to be observed, can be to a certain extent inferred from the theory developed. As demonstrated above the organismal properties that favour the process are short binding site length (parameter l in the model), high interference (large values of parameter B) and low average energetic costs of a mismatch (parameter b), while the population genetic properties that favour the process are lower effective population sizes (N_e), and low selective costs for overexpression (σ_2). These conditions are more commonly fulfilled in Eukaryotic compared to Prokaryotic organisms. The average length of a TFBS in Eukaryotes is 10 nucleotides, while the average in Prokaryotes is 16 nucleotides (Stewart et al., 2012), Eukaryotes have higher levels of interference

(parameter B) (Lynch, 2024; Wunderlich and Mirny, 2009). Effective population sizes (N_e) differ between groups by 5 orders of magnitude in the living world, with some Prokaryotes going as high as 10^9 and some higher Eukaryotes hitting the lower end at 10^4 individuals (Lynch, 2007).

The selective costs for overexpression can be hard to determine in a general way. Overexpression can lead to ectopic expression in multicellular organisms leading to metabolic and developmental irregularity (Farley et al., 2015). The selective effects of these changes can vary between organism and genes in question. However, the energetic costs of overexpression can be subject to a more detailed analysis (Lynch and Marinov, 2015; Wagner, 2005). In particular, Lynch and Marinov (2015) show that in Prokaryotes due their energetic capacity and high effective population sizes, even a duplication event of 10 nucleotides can be efficiently purged from the population just due to unnecessary higher energetic costs of DNA replication. On the other hand, for many higher Eukaryotes and especially higher Metazoans, unnecessary increases in the transcription and translation rates can still fall under effective neutrality, or be under weak selection (Lynch and Marinov, 2015).

All theses biological considerations imply that functional diffusion is a process expected to be more often realised in Eukaryotic species with lower effective population sizes and to be much more rare in Prokaryotes.

5.2 Functional diffusion and evolvability

Usually defined as "the ability of a biological system to produce phenotypic variation that is both heritable and adaptive" (Payne and Wagner, 2019), the notion of evolvability has been under much debate in the literature in recent decades (Pigliucci, 2008; Hansen et al., 2023). In particular the question of whether evolvability evolves as product of natural selection promoting it or as byproduct of selection (or other mechanisms) is still debated. The process of functional diffusion outlined here shows how evolvability can arise as a byproduct of suboptimization. The shift from a single-site state to a two-site state has enhanced the biological system's ability for regulatory co-option. The increase in evolvability is a byproduct of selection not being able to maintain the optimal design. The novel duplicated binding site (or the initial binding site) can act as pre-sites for the the evolution of novel TF binding sites that bind different transcription factor (TF) proteins. Given that TF proteins evolved through duplication and divergence events (Friedlander et al., 2017) the consensus binding sequence between two binding sites of different TF proteins is more similar than expected by chance (Tuğrul et al., 2015).

5.3 Evolution on the edge of the drift barrier and the origins of molecular complexity

The ability of selection to maintain a biological property, to promote advantageous mutations or to purge deleterious mutations is determined by the relation $2 \cdot N_e \cdot s >> 1$. When the values of this product are much larger than one, then selection is able to highly optimise the system, as long as beneficial mutations are mutationally available. When the product is $<< 1$, then selection is not strong enough to maintain the optimal design, and the system evolves under neutral dynamics (effective neutrality). The drift barrier hypothesis states that selection will optimise the system until it reaches a point where all further adaptive mutations have too little of a benefit to be optimised further (Lynch, 2024). This can be seen in Figure 2, as the product of $N_e \cdot a$ gets larger so does the distribution approach more optimal states. It has been argued that populations that evolve under lower N_e accumulate more slightly deleterious mutations and that this process has led to the origin of complex genomic architecture in Eukaryotes (Lynch, 2007). A more complex view of the evolution of molecular complexity has also been developed by Lynch (2012), under which large scale fitness increases happen by the emergence of novel cellular features that improve on previously established mechanisms, but are later sub-optimised due to slightly deleterious mutations. A layering effect can arise from this, where a semi-ratchet mechanism could come into action. When a complex novel adaptive mechanism arises, the fitness increase is then only transient, since more fine-grade slightly deleterious mutations degrade the overall functionality of the system to the previous (pre-adapted) state. Given the fact that the population is now again in the same fitness category, a large-scale adaptation is again possible and later slightly deleterious mutations can again accumulate. In such a view, rather counter-intuitively, smaller populations with sub-optimised features can be more prone to both adaptive and neutral evolution, the former since they can more easily explore different adaptive peaks on a fitness landscape and the later since more mutations fall under the effectively neutral category. While the mathematical framework developed by Lynch (2012) is more general, a particular application of that line of thinking can be seen in the functional diffusion model.

The origins of organismal complexity is a subject of a long standing debate (Lynch, 2007; Stolzfus, 1999). The theory that complexity comes about as a byproduct of lower lower effective population sizes allowing more deleterious mutational events to fix has gained considerable support since its first proposal (Lynch, 2007; Lynch, 2024; Bingham and Ratcliff, 2024). The theory developed here goes to show that this approach should be taken into account when considering the evolution of gene regulation.

5.4 Conclusion and outlook

If the model outlined above is true, some general and rather interesting implications follow. Firstly, the evolution of complex regulation, similar to the origins of complex genomic architecture and cellular features (Lynch, 2007; Lynch, 2024), is at least partly a product not of any particular environmental selection preferring such complex adaptations, but rather a shift to particular population genetic conditions. Secondly, sub-optimisation that comes about as a consequence of selection being unable to maintain the optimal form can aid evolvability and complex adaptations by opening mutational pathways that are closed off to populations that are well maintained at an adaptive peak. Thirdly, selection can promote the origin of new gene duplicates without necessarily having neofunctionaliztion. The functional diffusion model when interpreted as a model of gene duplication rather then TFBS duplication lies in between models of subfunctionaliztion and neofunctionalization (the former since it leads initially to no novel function and the latter since it has selection favouring the duplication event). Finally, the model outlined here has the potential to partially explain to what Nick Barton has refereed to as diffusive function in the genome (Barton, 2022), the observation that in Eukaryotes genomic function is not localised but diffused between several binding sites and genes.

Further theoretical research on this topic needs to consider whether multiple consecutive duplication and degeneration events are possible leading to a ratchet mechanism and an origin of clusters of weak binding sites that are often found in Eukaryotes (Nourmohammad and Lässig, 2011; Wunderlich and Mirny, 2009). Integrating the theory presented can with current molecular population genetics can aid the effort to see how often functional diffusion happens and how it is distinguished from other processes.

6 Methodology

In order to visualise the theoretical predictions of the analytical expressions used and developed here programs R (version 4.3.1) and Phyton (Jupiter notebook version 6.5.4) were used. Wright-Fisher simulations were done using Phyton (Jupiter notebook version 6.5.4). Large language models (ChatGPT 3.5) were used for code debugging and for improving and modifying the code. The code for the simulations and analytical predictions for all the graphs developed here are provided in the Supplementary material.

7 References

1. Barton, Nicholas H. "The "new synthesis"." Proceedings of the National Academy of Sciences 119.30 (2022): e2122147119.

2. Bénitière, Florian, Anamaria Necsulea, and Laurent Duret. "Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans." Elife 13 (2024): RP93629.

3. Bingham, Emma P., and William C. Ratcliff. "A nonadaptive explanation for macroevolutionary patterns in the evolution of complex multicellularity." Proceedings of the National Academy of Sciences 121.7 (2024): e2319840121.

4. Carroll, Sean B. "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution." Cell 134.1 (2008): 25-36.

5. Eddy, Sean R. "The C-value paradox, junk DNA and ENCODE." Current biology 22.21 (2012): R898-R899.

6. Farley, Emma K., et al. "Suboptimization of developmental enhancers." Science 350.6258 (2015): 325-328.

7. Friedlander, Tamar, et al. "Evolution of new regulatory functions on biophysically realistic fitness landscapes." Nature communications 8.1 (2017): 216.

8. Gerland, Ulrich, and Terence Hwa. "On the selection and evolution of regulatory DNA motifs." Journal of molecular evolution 55 (2002): 386-400.

9. Hansen, Thomas F., David Houle, Mihaela Pavlicev, and Christophe Pélabon, eds. Evolvability: A unifying concept in evolutionary biology?. MIT Press, 2023.

10. He, Xionglei, and Jianzhi Zhang. "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution." Genetics 169.2 (2005): 1157-1164.

11. Hermisson, Joachim, and Pleuni S. Pennings. "Soft sweeps: molecular population genetics of adaptation from standing genetic variation." Genetics 169.4 (2005): 2335-2352.

12. Keren, Leeat, et al. ”Massively parallel interrogation of the effects of gene expression levels on fitness.” Cell 166.5 (2016): 1282-1294.

13. Kimura, Motoo. ”On the probability of fixation of mutant genes in a population.” Genetics 47.6 (1962): 713.

14. Koonin, Eugene V. ”Darwinian evolution in the light of genomics.” Nucleic acids research 37.4 (2009): 1011-1034.

15. Koonin, Eugene V. ”Evolution of genome architecture.” The international journal of biochemistry and cell biology 41.2 (2009): 298-306.

16. Lässig, Michael. ”From biophysics to evolutionary genetics: statistical aspects of gene regulation.” BMC bioinformatics (2007): S7.

17. Lynch, Michael. ”Evolutionary layering and the limits to cellular perfection.” Proceedings of the National Academy of Sciences 109.46 (2012): 18851-18856.

18. Lynch, Michael. ”The frailty of adaptive hypotheses for the origins of organismal complexity.” Proceedings of the National Academy of Sciences 104.suppl_1 (2007): 8597-8604.

19. Lynch, Michael. ”Evolutionary layering and the limits to cellular perfection.” Proceedings of the National Academy of Sciences 109.46 (2012): 18851-18856.

20. Lynch, Michael. ”The origins of genome architecture.” Vol. 98. Sunderland, MA: Sinauer associates, 2007.

21. Lynch, Michael R. Evolutionary cell biology: the origins of cellular architecture. Oxford University Press, 2024.

22. Lynch, Michael, and Bogi Trickovic. ”A theoretical framework for evolutionary cell biology.” Journal of molecular biology 432.7 (2020): 1861-1879.

23. Lynch, Michael, and John S. Conery. ”The origins of genome complexity.” science 302.5649 (2003): 1401-1404.

24. Lynch, Michael, and Kyle Hagner. ”Evolutionary meandering of intermolecular interactions along the drift barrier.” Proceedings of the National Academy of Sciences 112.1 (2015): E30-E38.

25. Lynch, Michael, and Georgi K. Marinov. ”The bioenergetic costs of a gene.” Proceedings of the National Academy of Sciences 112.51 (2015): 15690-15695.

26. Mustonen, Ville, et al. ”Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites.” Proceedings of the National Academy of Sciences 105.34 (2008): 12376-12381.

27. Nourmohammad, Armita, and Michael Lässig. ”Formation of regulatory modules by local sequence duplication.” PLoS Computational Biology 7.10 (2011): e1002167.

28. Ohta, Tomoko. ”Slightly deleterious mutant substitutions in evolution.” Nature 246.5428 (1973): 96-98.

29. Payne, Joshua L., and Andreas Wagner. ”The causes of evolvability and their evolution.” Nature Reviews Genetics 20.1 (2019): 24-38.

30. Pigliucci, Massimo. ”Is evolvability evolvable?.” Nature Reviews Genetics 9.1 (2008): 75-82.

31. Rice, Sean. ”Evolutionary theory: mathematical and conceptual foundations.” (2004).

32. Sandve, Simen R., Rori V. Rohlfs, and Torgeir R. Hvidsten. ”Subfunctionalization versus neofunctionalization after whole-genome duplication.” Nature genetics 50.7 (2018): 908-909.

33. Sella, Guy, and Aaron E. Hirsh. ”The application of statistical physics to evolutionary biology.” Proceedings of the National Academy of Sciences 102.27 (2005): 9541-9546.

34. Stewart, Alexander J., Sridhar Hannenhalli, and Joshua B. Plotkin. ”Why transcription factor binding sites are ten nucleotides long.” Genetics 192.3 (2012): 973-985.

35. Stoltzfus, Arlin. ”On the possibility of constructive neutral evolution.” Journal of molecular evolution 49 (1999): 169-181.

36. Tuğrul, Murat, et al. ”Dynamics of transcription factor binding site evolution.” PLoS genetics 11.11 (2015): e1005639.

37. Wagner, Andreas. ”Energy constraints on the evolution of gene expression.” Molecular biology and evolution 22.6 (2005): 1365-1374.

38. Walsh, Bruce, and Michael Lynch. Evolution and selection of quantitative traits. Oxford University Press, 2018.

39. Wunderlich, Zeba, and Leonid A. Mirny. ”Different gene regulation strategies revealed by analysis of binding motifs.” Trends in genetics 25.10 (2009): 434-440.