Adaptive gene misregulation

Andreas Wagner 1,2,3,*

1Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, CH-8057, Switzerland
2The Santa Fe Institute, Santa Fe, NM 87501, USA
3Swiss Institute of Bioinformatics, Lausanne, Switzerland

*Address for correspondence: andreas.wagner@ieu.uzh.ch

Abstract

Because gene expression is important for evolutionary adaptation, its misregulation is an important cause of maladaptation. A misregulated gene can be incorrectly silent ("off") when a transcription factor (TF) that is required for its activation does not bind its regulatory region. Conversely, a misregulated gene can be incorrectly active ("on") when a TF not normally involved in its activation binds its regulatory region, a phenomenon also known as regulatory crosstalk. DNA mutations that destroy or create TF binding sites on DNA are an important source of misregulation and crosstalk. Although misregulation reduces fitness in an environment to which an organism is well-adapted, it may become adaptive in a new environment. Here, I derive simple yet general mathematical expressions that delimit the conditions under which misregulation can be adaptive. These expressions depend on the strength of selection against misregulation, on the fraction of DNA sequence space filled with TF binding sites, and on the fraction of genes that must be expressed for optimal adaptation. I then use empirical data from RNA sequencing, protein-binding microarrays, and genome evolution, together with population genetic simulations to ask when these conditions are likely to be met. I show that they can be met under realistic circumstances, but these circumstances may vary among organisms and environments. My analysis provides a framework in which improved theory and data collection can help us demonstrate the role of misregulation in adaptation. It also shows that misregulation, like DNA mutation, is one of life's many imperfections that can help propel Darwinian evolution.

Keywords: evolvability; adaptation; regulation; expression

Introduction

A typical eukaryotic genome encodes more than 10,000 genes whose expression is regulated by the binding of specialized proteins called transcriptional regulators (transcription factors, TFs) to regulatory DNA near a gene. A TF usually binds a short sequence motif (TF binding site) that helps activate or repress the gene's transcription. An important source of gene misregulation is DNA mutations that either destroy TF binding sites or create such sites in inappropriate places, possibly from "presites" that are similar to TF binding sites (MacArthur and Brookfield 2004; Tuğrul et al. 2015). Two fundamental kinds of misregulation can be distinguished. First, a gene is wrongly inactive ("off"). It is not expressed where or when it should be expressed, for example, because a mutation has destroyed the nearby binding site of an activator, or because a mutation has created a binding site for a repressor. Second, a gene is wrongly active ("on"). That is, the gene is expressed where or when it should not be expressed, for example because a mutation has created a new TF binding site for an activator or destroyed a binding site for a repressor. The de novo creation of TF binding sites by mutation can be frequent when TF binding sites are abundant in DNA sequence space (Stewart et al. 2012; Tuğrul et al. 2015). The resulting new and undesirable gene regulation is also known as regulatory crosstalk (Wunderlich and Mirny 2009; Friedlander et al. 2016). Such crosstalk can even occur in the absence of mutations, because many TFs bind multiple sites on DNA and can activate or repress transcription from noncognate sites. Regulatory crosstalk belongs to a class of broader phenomena in cell-biology, which can also involve maladaptive interactions between transmembrane receptors, protein kinases and protein phosphatases (Hill 1998; Danielpouir and Song 2006; McClean et al. 2007; Junttila et al. 2008; Rowland et al. 2012, 2017; McClune et al. 2019; McClune and Laub 2020).

Misregulation and crosstalk are ubiquitous and inevitable consequences of large genomes and regulatory complexity (Wunderlich and Mirny 2009; Friedlander et al. 2016). During the evolution of eukaryotes, genomes have increased in size, number of genes, and in the proportion of noncoding and thus potentially regulatory DNA (Lynch and Conery 2003; Lynch 2007). In addition, the number of TFs encoded in a genome has increased faster than linearly with gene number (van Nimwegen 2006). The potential for misregulation and crosstalk has increased concomitantly. They are systemic properties whose incidence cannot be understood from considering one or few genes, but only from studying gene regulation genome-wide. According to one recent estimate, regulatory crosstalk may affect as many as 23% of eukaryotic genes (Friedlander et al. 2016).

The expression of a specific gene may be deleterious in one environment but beneficial in another, and the same holds for a
gene that is not expressed in any one environment. This raises the possibility that misregulation that is deleterious in one environment may help create novel gene expression patterns that are adaptive in a new environment. Such “adaptive misregulation” may play an important role when genes evolve de novo from non-protein coding DNA, because such genes often originate from newly transcribed noncoding DNA (Begun et al. 2007; Zhao et al. 2014; Ruiz-Orera et al. 2015; Neme and Tautz 2016; Majic and Payne 2020).

Here, I aim to identify conditions under which misregulation can become beneficial in a new environment colonized by an organism or in a new cellular state, such as a new cell type (Arendt et al. 2016; Wagner et al. 2019). Under these conditions, misregulation would be beneficial for the same reasons that some DNA mutations are. Natural selection cannot reduce the rate of DNA mutations to zero (Lynch et al. 2016), even though most mutations are deleterious in environments to which an organism is well adapted (Eyre-Walker and Keightley 2007). In a new environment, some mutations can be fortuitously beneficial, and their benefits ultimately drive Darwinian evolution. Likewise, the benefits of misregulation would be fortuitous side-effects of natural selection’s inability to eliminate misregulation completely.

I begin by describing simple yet general conditions under which misregulation may be adaptive. I then ask whether these conditions can ever be met, given what we know about the abundance of TF binding sites in sequence space, and how many genes an organism typically expresses. The answer is yes, but it depends on factors that may vary across organisms, such as the organism’s ability to create and destroy binding sites with probabilities \( \mu^+ \) and \( \mu^- \), respectively. I estimated these mutation probabilities for a given value of the fraction of sequence space \( p_b \) filled with binding sites from high throughput data on mouse TF-DNA binding data, as I explain in a separate section of the Materials and Methods.

In each generation, I determined the number of binding sites to be destroyed among the \( n_{11} \) binding sites as \( d_{11} = P(\mu^+ n_{11}) \), which denotes a Poisson-distributed pseudorandom variate with mean \( \mu n_{11} \). Likewise, I determined the number of new binding sites to be created from the \( n_{00} \) inactive binding sites by a pseudo-random variate \( c_0 = P(\mu^- n_{00}) \). I then determined the number \( n'_{11} \) of active binding sites after mutation as \( n'_{11} = n_{11} - d_{11} + c_{10} \), and the number \( n'_{10} \) of inactive binding sites after mutation as \( n'_{10} = n_{10} + d_{11} - c_{10} \). Very rarely this procedure might yield a negative number of binding sites after mutation. In this case, if \( n'_{11} < 0 \), I set \( n'_{11} = 0 \) and \( n'_{10} = n_{10} \). Conversely, if \( n'_{10} < 0 \), I set \( n'_{10} = 0 \) and \( n'_{11} = n_{11} \). This ensures that mutation preserves the required relationship \( n'_{11} + n'_{10} = n_{11} \). I proceeded analogously for the binding sites associated with genes that are not to be expressed, computing \( n'_{01} = n_{01} - d_{01} + c_{00} \), as well as \( n'_{00} = n_{00} + d_{01} - c_{00} \), where \( d_{01} \) is a \( P(\mu^- n_{01}) \) pseudo-random variate, and \( c_{00} \) is a \( P(\mu^- n_{00}) \) pseudo-random variate. I note that my model represents neither the identity of TFs nor their number or expression directly, but only indirectly through the fraction \( p_b \) of sequence space that constitutes active TF binding sites.

After mutation, I computed the fitness \( w_i \) for each individual \( i \) in the population as \( w_i = (1 - s_{10})^{\frac{n_i}{G}}(1 - s_{01})^{\frac{n_i}{G^*}} \), where \( s_{01} \) and \( s_{10} \) denote the number of incorrectly off genes (inactive binding sites) and incorrectly on genes (active binding sites) of individual \( i \) after mutation, respectively. I then normalized (divided) each fitness value by the maximal fitness \( w_{\max} = w_{ij} / w_{\max} \) in the population. To implement (soft) selection, I repeated the following procedure until a post-selection population of \( N_e \) individuals had been created. I chose one individual \( i \) at random from the population, and created a pseudo-random variate \( v \) that is uniformly distributed on the interval \( (0, 1) \). If \( v < w_{\min} \), I placed the individual in the post-selection population, otherwise I rejected the individual, and repeated the process with another individual chosen at random in the same way (with replacement). I continued sampling individuals in this manner until I had created a new population of \( N_e \) (surviving individuals). I implemented these simulations in python 3.7 with the numpy library for numerical analysis.

Sensible ranges for some of the parameters in the model, I use here can be estimated from available functional genomic data. The next three sections explain how to estimate these parameter ranges.

### Materials and methods

#### Population genetic simulations

Here, I explain how I simulated the evolutionary dynamics of the misregulation model I develop in Results. I used a discrete-time Wright-Fisher model (Hartl and Clark 2007) in which cycles of mutation and viability selection alternate in a population with \( g(\mu) \). The model represents each individual by a number of genes \( G \), of which \( G_n \) genes should be expressed under optimal adaptation and the remaining \( G_{off} \) genes should not be expressed \( (G_n + G_{off} = G) \). The model considers only transcriptional activation, and I assume that the expression state of any one gene is determined by a single TF binding site associated with the gene. This site may be a high-affinity (“active”) binding site for a TF, in which case I assume that the gene is expressed. It may also be a low (or no) affinity (“inactive”) binding site, in which case the gene is not expressed.

An important simplifying assumption is that recombination is absent, that is, all TF binding sites behave as if they occurred on a single nonrecombining chromosome. For computational feasibility, the model does not represent binding sites individually, but instead represents the total number of (i) active binding sites for genes that should be active \( (n_{11}) \), (ii) inactive binding sites for genes that should be active \( (n_{10}) \), (iii) active binding sites for genes that should be inactive \( (n_{01}) \), and (iv) inactive binding sites for genes that should be inactive \( (n_{00}) \). The fractions of these binding sites, and hence the fraction of active/inactive genes computes as \( f_{11} = n_{11}/G \), e.g. \( f_{10} = n_{10}/G \). Where useful or necessary, I sometimes also use an alternative normalization, where \( f_{11} = n_{11}/G_{on} \), and \( f_{10} = n_{10}/G - G_{on} \), and mention this alternative normalization.

At the beginning of each population simulation, I initialized all individuals \( i \) such that fitness \( w_i = 1 \), i.e. by setting \( n_{11} = G_{on} \) and \( n_{00} = G_{off} \). Mutations in binding sites of genes that should be active create and destroy binding sites with probabilities \( \mu^+ \) and \( \mu^- \), respectively. I estimated these mutation probabilities for a given value of the fraction of sequence space \( p_b \) filled with binding sites from high throughput data on mouse TF-DNA binding data, as I explain in a separate section of the Materials and Methods.

### Estimating the fraction of expressed genes and gene expression correlations c from human and mouse tissue-specific RNA sequencing data

I downloaded previously published RNA sequencing (RNA seq) data from human and mouse (Cardoso-Moreira et al. 2019) in the form of reads per kilobase pair of transcript per million mapped reads (rpkm) from http://evodevoapp.kaessmannlab.org (March 2020).
These data catalogs mRNA expression for multiple tissues and in multiple developmental stages (including adults) of multiple mammals. For mouse I used data from the following seven tissue samples obtained 28 days post-partum: Brain P28.1, Cerebellum P28.1, Heart P28.1, Kidney P28.1, Liver P28.1, Ovary P28.1, Testis P28.1. For humans, I used the following data sets, which stem from either young adults (5 tissues), from an 8-year-old child (kidney), or from an 18-week-old fetus [ovaries, 18 weeks post conception (wpc)]: Brain youngAdult 47, Cerebellum youngAdult 51, Heart young Adult 49, Kidney school 40, Liver youngAdult 45, Ovary 18wpc 18, and Testis youngAdult 36. I then obtained a complete set of ensemble gene identifiers (ENSG IDs) for all human and mouse protein coding genes from biomart (http://www.ensembl.org/biomart, March 2020), and extracted from each of the above RNA seq data sets the data corresponding to human or mouse protein coding genes. I then determined the maximal expression level for all tissues of the same organism, and computed the fraction of genes in each sample whose expression exceeds a proportion $P = 10^{-5}$ of this maximal expression level. I chose this interval, because mRNA expression levels range over approximately five orders of magnitude (Wang et al. 2019). I then also determined the overlap in the fraction of expressed genes for each pair of tissues in the same organism, by determining the fraction of genes that were expressed in both tissues, and compared it to the expected fraction if the gene expression states were stochastically independent. If $f_a$ and $f_b$ is the fraction of genes expressed in tissue $a$ and $b$, then this expected fraction calculates as $f_a f_b$.

### Estimating the fraction $p_b$ of sequence space filled with binding sites from protein binding microarray data

My source of DNA binding data and other information for individual TFs was the database CIS-BP, which hosts a compendium of DNA binding data from thousands of TFs and many species (Weirauch et al. 2014). I obtained such data for all TFs in six model organisms. These are the yeast Saccharomyces cerevisiae, the round worm Caenorhabditis elegans, the fruit fly Drosophila melanogaster, the thale cress Arabidopsis thaliana, the mouse Mus musculus, and Homo sapiens. I downloaded bulk data from http://cisbp.ccb.rutgers.edu/cisbp.php (March 2020), which included a file containing general information for each TF (file “TF_information_all_motifs”), protein binding microarray (PBM) data (file “Escores.txt”), and position weight matrix data (Benos et al. 2002) that indicate the contribution of each nucleotide position in a binding site to binding affinity (directory “pwms_all_motifs”, Weirauch et al. 2014). For TFs with more than one associated PBM experiment or PWM, I chose an arbitrary experiment and PWM. I calculated the length distribution of TF binding sites in any one species from the position weight matrices provided for each TF by CISBP (directory “pwms_all_motifs”).

The TF-DNA binding data, I use comes from PBMs (Berger et al. 2006, Badis et al. 2009). Briefly, a PBM harbors all double-stranded oligonucleotides of length $L = 10$, and allows one to measure how strongly a given TF binds to each oligonucleotide (Berger et al. 2006, Badis et al. 2009, Weirauch et al. 2014). TF binding data are usually reported for eight-mers, because each eight-mer is contained multiple times in the ten-mers on an array, which allows for more reliable protein binding measurements through the resulting measurement replication. There are $(4^8 - 4^4)/2 = 4^4 = 32,896$ eight-mers, a number that is smaller than $4^8 = 65,536$, because TF binding does not distinguish between an eight-mer and its reverse complement, and because $4^4$ eight-mers are palindromes, and thus identical to their reverse complement. PBM data are usually reported in the form of enrichment (E-)scores, which are reproducible rank-ordered array signal intensity values for different eight-mers that lie in the interval $(-0.5,0.5)$. They are robust to changes in TF concentration, allow comparison across TFs, correlate with binding affinity, and can thus be thought of as relative affinity values (Berger et al. 2006; Badis et al. 2009; Payne and Wagner 2014).

My analysis is based on binding sites with an E-score exceeding 0.35, because such high E-scores reflect high-affinity binding, and are associated with a low-discovery rate (Berger et al. 2006; Badis et al. 2009; Nakagawa et al. 2013; Payne and Wagner 2014). The number of such high-affinity sites varies across species and TFs. It ranges between 24 and 2323 sites per TF in yeast (PBM data available for 56 TFs), between 21 and 2509 sites in Arabidopsis (167 TFs), between 180 and 1822 sites in the fruit fly (20 TFs), between one and 2030 sites in the round worm (112 TFs), between 61 and 2436 sites in the mouse (144 TFs), and between 15 and 2254 sites in humans (57 TFs) (Weirauch et al. 2014).

For each one of the six species listed above, I established a list of all TFs according to their CIS-BP TF identifier (CIS-BP ID), and extracted from this list those TFs whose binding sites did not exceed eight base pairs in length. They comprise the following fractions of all TFs for which CIS-BP contains data for any one species: 28% (yeast), 16% (Arabidopsis), 18% (fruit fly), 40% (round worm), 15% (mouse), and 12% (human). For any one species, and for each TF $i$ in this species for which PBM data was available, I determined the set of all binding sites $B_i$ with an E-score exceeding 0.35. For each nonpalindromic site I added the reverse complement of the site to this set. Subsequently, I randomly shuffled the order of TFs in the resulting data set for each species. I then computed the cumulative union of these sets, i.e. $B_n = \bigcup_{i=1}^n B_i$, where I allowed the index $n$ to range from one to all $n$ TFs for which PBM data are available in a given species. For any given $n$, the quantity $p_n = |B_n|/4^8$ is the fraction of sequence space ($L=8$) occupied by TF binding sites.

To integrate the PBM data with mRNA expression data from humans and mice, I used the same RNA seq data I described above (Cardoso-Moreira et al. 2019), and restricted my analysis of PBM data to those TFs whose expression was above a fraction $10^{-5}$ of the maximal expression level for all seven tissues. To integrate PBM data with proteome-based human gene expression data, I used the human protein atlas (Uhlen et al. 2015) (version 19.3, ensemble version 92.38, https://www.proteinatlas.org/about/download, file “normal_tissue.tsv”, March 2020), which catalogs the expression of 74% (15,313/20,344) of human protein coding genes in 44 normal organs or tissue types based on immunohistochemistry (see also https://www.proteinatlas.org/human/proteome/tissue). For data from each tissue type, I identified ensemble gene identifiers for those genes whose expression level was not recorded as “Not Detected” and for which the reliability of the expression measurement was not recorded as “Uncertain.” I then restricted my above analysis of PBM data to those TFs whose ensemble gene identifiers had a representative in the resulting expression data set.

### Estimating mutation probabilities for random and biological TF binding sites

I first numerically estimated the probability of creating or destroying binding sites under the assumption that some fraction $p_b$ of a space of DNA sequences of length $L$ constitutes high-affinity binding sites that are randomly and uniformly distributed in this space. Although this random model is unrealistically simple, it serves as a useful point of comparison for biological
binding sites. Specifically, I first created for a given L a list of all DNA sequences with length L. Then, I randomly permuted this list, and chose the first $p_0 4^L$ entries as a hypothetical set B of binding sites. I subsequently examined each sequence S in the space, and created a single random mutation in S by randomly altering one of its L nucleotides (chosen at random with uniform probability) to one of three possible alternative nucleotides. If the sequence S was in the set B, I recorded whether the mutation created a sequence that was not in B (i.e. not a binding site). Likewise, if S was not in B, I recorded whether the mutation did create a sequence in the set B. In addition, I recorded whether S was in the 1-neighborhood of the set of B binding sites. I repeated this procedure (random permutation, selection of $p_0$ binding sites, and mutating each binding site) $n = 10$ times, and computed the probability of binding site creation and destruction from these ten replicates. Error bars in Supplementary Figure S2 correspond to one standard error of the mean over these replicates.

I used the same procedure with $n = 3$ replicates to estimate the probability that biological binding sites become created or destroyed, except that I used not a random sample of binding sites that fills a fraction $p_0$ of sequence space of size $4^L$, but a set of binding sites derived from mouse PBM data that is as close as possible to a given value of $p_0$. Specifically, I computed for a given value of $p_0$, as described above, the cumulative union $B_n$ of sets of binding sites $B_i$ for mouse TFs i, i.e. $B_n = \cup_{i=1}^n B_i$, where the index $n$ is the smallest $n$ such that $|B_n|/4^L > p_0$. I used mouse PBM data for this purpose, because of the large number of TFs (144) whose binding sites of length $L \leq 8$ have been identified with PBMs.

Data availability
All experimental data that I have used in this analysis is publicly available gene expression and PBM data. Simulation code is available on github at https://github.com/andreas-wagner-uzh/misregulation-simulations.

Supplementary material is available at figshare at https://doi.org/10.25386/genetics.13370144.

Results

General conditions under which misregulation can be beneficial
I consider two different states that an organism can be in, an evolutionarily ancestral one and a derived one. The ancestral state could be elicited by an environment that the organismal lineage often encounters, or by a pathogen to which the organismal lineage is frequently exposed. Alternatively, one can think of it as a well-established cell or tissue type that is vital to the organism’s life cycle. For brevity, I will simply refer to it as the “old” state, and assume that the organism’s gene expression pattern is well-adapted to it, within the limits imposed by inevitable misregulation. I assume that the derived state is evolutionarily so new that the organism has never encountered it and is therefore ill-adapted to it. This new state might be elicited by a novel abiotic environment that an organism is colonizing and that may harbor, for example, an anthropogenic toxin or an emergent pathogen to which the organism is initially not well adapted. A new cell or tissue type might help the organism survive in this new environment. I envision that the organism occupies the new state in permanence, akin to a species invading a new habitat, and need not cycle back and forth between the old and the new state. I assume that poor adaptation to the new state is reflected in a gene expression pattern that is unchanged relative to the old state. I will ask under which conditions misregulation in the old state may alleviate the poor adaptation to the new state.

Like some previous studies on misregulation and crosstalk (Friedlander et al. 2016, Grah and Friedlander 2020), my model considers gene expression a binary variable and only distinguishes between genes that are expressed (“on,” “active,” “1”) or not expressed (“off,” “inactive,” “0”). I subdivide all G genes of an organism into those $G_{on}$ genes that should be expressed in the old state if the organism were optimally adapted to this state, and those $G - G_{on}$ genes that should be off. I denote the fraction of genes that should be on in the old state as $f^{o} = G_{on}/G$. Because of misregulation, the actual fraction of genes $f^{\pm}_G$ that are on is not necessarily equal to $f^{o}$, and the fraction of genes that are off $f^{\pm}_G = 1 - f^{\pm}_G$ is not necessarily equal to $1 - f^{o}$. It is useful to subdivide the misexpressed genes into a fraction $f^{\pm}_{G_{on}}$ of such genes that are active, although for optimal adaptation they should not be active (wrongly on), and a fraction $f^{\pm}_{G_{off}}$ of genes that are inactive although they should be active (wrongly off). Analogously, I will denote the fraction of genes that are correctly active and correctly inactive as $f^{10}_s$ and $f^{01}_s$, respectively. With this notation, the following identities hold as,

$$f^{o} = f^{10}_G + f^{01}_G$$

$$f^{\pm}_G = f^{10}_G + f^{01}_G$$

$$1 - f^{\pm}_G = f^{01}_G + f^{10}_G$$

I quantify the total extent of misregulation in the old state by the total fraction of misexpressed genes, i.e. by $f^{10}_G + f^{01}_G$. I note that $f^{10}_G + f^{01}_G + f^{10}_G + f^{01}_G = 1$. I assume that selection acts on viability, that each misexpressed gene reduces fitness independently from other genes, and that these fitness reductions contribute multiplicatively to organismal fitness. More specifically, I assume that each wrongly off gene reduces fitness by a factor $1 - s_{10}$, and that each wrongly on gene reduces fitness by a factor $1 - s_{01}$, where s denotes a selection coefficient associated with gene misexpression. Overall, the fitness of an individual is then given by $w = (1 - s_{10} f^{10}_G)(1 - s_{01} f^{01}_G)$. The selection coefficients $s_{10}$ and $s_{01}$ are typically very small (Hahn et al. 2003; Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009), such that fitness may be well approximated by $w \approx 1 - G(s_{10} f^{10}_G + s_{01} f^{01}_G)$, which assumes that terms involving $s_{10} s_{10}$ can be neglected. The right-most term represents the reduction in fitness caused by misregulation.

My model represents poor adaptation of an organism to the new state by unchanging gene expression relative to the old state. In contrast, if an organism was optimally adapted to the new state, it would express some fraction $f^{0}_G$ of its genes in this state, which may be different from the genes expressed in the old state. To study the relationship between $f^{0}_G$ and the genes expressed in the old state, I will treat the expression states of individual genes as random variables. Specifically, I define binary random variables $g_i^+$ for all G genes, such that $g_i^+ = 1$ if gene $g_i$ is actually expressed, and $g_i^+ = 0$ if it is not expressed in the old state. In consequence, $F(g_i^+ = 1) = f^{0}_G$. I define a similar variable $g_i^-$ for optimal expression in the new state, such that $F(g_i^- = 1) = f^{0}_G$. Empirical data suggests that these variables are generally positively correlated (File S7, Supplementary Figure S1, B and C show examples from humans and mouse.). In other words, it is more likely that a gene already expressed in the old state also needs to be expressed in the new state than expected by chance alone. I will quantify this correlation by a variable c, which I allow to range from $c = 0$ for no correlation to $c = 1$ for a perfect correlation, where any gene that is (not) expressed in the old state needs (not) be
expressed in the new state. One can show (Supplementary File S6) that $c$ is identical to a Pearson correlation coefficient, except for a linear scaling factor, whenever gene expression states are positively correlated.

My main goal is to find out whether the fraction of correctly expressed genes and/or fitness in the new state can increase when misregulation is present compared to when it is absent. To answer this question, one needs to know several quantities. These include the fraction of genes that are (i) correctly active in the new state ($f_{N01}$), (ii) correctly inactive in the new state ($f_{N00}$), (iii) incorrectly active in the new state ($f_{M01}$), and (iv) incorrectly inactive in the new state ($f_{M00}$). I show how to calculate these and other useful quantities in File SS, and summarize my calculations here. I will first consider the change in the fraction of correctly expressed genes in the new state, when misregulation is present in the old state, as opposed to when it is absent ($f_{N01} = 0, f_{N00} = 0$). To this end, I define the quantity $D_{fN} = f_{N01} - f_{N00}$, where the right-most term means that the fraction $f_{N01}$ of correctly active genes in the new state is evaluated in the absence of misregulation in the old state. This quantity is the difference between the fraction of correctly active genes in the new state if misregulation is present in the old state, as opposed to when it is absent ($f_{N01} = 0, f_{N00} = 0$). To this end, I define the quantity $D_{fN} = f_{N01} - f_{N00}$, where the right-most term means that the fraction $f_{N01}$ of correctly active genes in the new state is evaluated in the absence of misregulation in the old state. This quantity is the difference between the fraction of correctly active genes in the new state if misregulation is present in the old state ($f_{N01}$) and if it is absent in the old state ($f_{N10}$). It is positive if misregulation increases the fraction of correctly active genes in the new state. I show in Supplementary File SS that this difference has the simple form:

$$\Delta_f^{m0} = f_{N01} - f_{N00} = c(1 - f^N)\].$$

(2)

Here, $D_{fN} = f_{N01} - f_{N00}$ designates a key quantity that I will call the excess of wrongly active genes under misregulation in the old state. It is positive if there are more wrongly active genes than wrongly inactive genes. Equation (2) implies that misregulation will increase the fraction of genes that are correctly active in the new state if $\Delta_m > 0$. This is because the second factor on the right-hand side is never negative. If $\Delta_m > 0$, the advantage of misregulation grows with an increasing expression correlation $c$ between the old and the new state. It also grows with an increasing fraction of genes that need to be expressed in the new state.

Analogous to $\Delta_f^{m0}$, I define $\Delta_f^{M0} = f_{N00} - f_{N00}$, as the difference in the fraction of genes that are correctly inactive when misregulation is present ($f_{N00}$) and when it is absent ($f_{N00}$). This quantity computes as (Supplementary File S5):

$$\Delta_f^{M0} = -\Delta_m(1 - c)(1 - f^N).$$

(3)

To compute the overall change in the fraction of correctly expressed genes, one can add Equations (2) and (3) to obtain the expression (Supplementary File S5):

$$\Delta_f^{m0} + \Delta_f^{M0} = \Delta_m[2f^N - 1](1 - c) + c].$$

(4)

Whether misregulation can increase the fraction of correctly expressed genes depends once again critically on the sign of $\Delta_m$. In addition, it now also depends on the sign and magnitude of $2f^N - 1$, because $c, 1 - c > 0$. For example, if there are more incorrectly active genes than incorrectly inactive genes ($\Delta_m > 0$) AND if more than half of all genes must be expressed in an organism well-adapted to the new state ($2f^N - 1 > 0$), then misregulation can lead to an increase in the fraction of correctly expressed genes in the new state.

Finally, one can use the same approach to determine how misregulation affects organismal fitness in the new environment. To this end, I define a ratio $r_w$ of fitness values with and without misregulation, which computes as (Supplementary File S5):

$$r_w := \frac{w^N}{w_0}\]$$

$$= (1 - s_{01})^{c_0}(1 - f^N) - (1 - s_{10})^{c_0}(1 - c) + c].$$

(5a)

$$= 1 + G_{10}[s_{01}(1 - c) + c] - s_{01}(1 - c)(1 - f^N)].$$

(5c)

where the approximation Equation (5c) is valid if $s_{01}$ and $s_{10}$ are small, and if terms involving $s_{01}^2, s_{10}$ can be neglected. If the ratio $r_w$ exceeds one, then misregulation provides a net fitness advantage, i.e. misregulation in the old state increases fitness in the new state. Again, $\Delta_m$ plays an important role in this ratio. Also, I note that differences in the strength of selection against incorrectly off genes ($s_{10}$) and incorrectly on genes ($s_{01}$) affect whether misregulation will increase or reduce fitness, and to what extent it will do so.

A population genetic model for the evolution of misregulation and its benefits

In the preceding section, I provided simple yet general conditions under which misregulation in an old organismal state can be beneficial, in the sense that it increases (i) the fraction of correctly active genes (2), (ii) the fraction of correctly inactive genes (3), (iii) the total fraction of correctly expressed genes (4), or (iv) organismal fitness (5). For the remainder of this contribution, I will ask when these conditions may be fulfilled in evolving populations that are subject to genetic drift, to natural selection, and to mutations that create and destroy the TF binding sites that are essential for activating genes.

Three quantities influence the adaptive benefit or burden of misregulation for a new cell state. The first is the excess $\Delta_m = f_{01} - f_{00}$ of wrongly active over wrongly inactive genes. The second is the fraction $f^0$ of genes that should be expressed in the new state under optimal adaptation. The third is the expression correlation $c$ between different expression states.

I will first turn to $\Delta_m$. Everything else being equal, a change in the sign of $\Delta_m$ turns misregulation from beneficial to detrimental, or vice versa. Since $\Delta_m$ and the incidence of misexpressed genes cannot be directly observed, it is necessary to predict their incidence from population genetic processes. To this end, I will consider a simple, haploid, asexual population genetic model for the evolution of gene regulation. The aim of this model is to estimate the fraction of wrongly active and inactive genes under broadly varying selection pressures in mutation-selection-drift equilibrium.

Like the calculations above, the model separates an organism’s genes into a subset of $G_{off}$ genes that should not be expressed for optimal adaptation, and a subset of $G_{on}$ genes that should be expressed $(1 - f^0 = G_{off}/G, f^0 = G_{on}/G)$. A major simplifying assumption is that the expression of each gene is determined by the binding of a transcriptional activator to a single high-affinity binding site near the gene. Experimental data from many transcriptional regulators in multiple species demonstrate that any one regulator can bind dozens to thousands of different sites with high affinity and specificity (Berger et al. 2006; Badis et al. 2009; Weirauch et al. 2014). Also, any one cell or tissue expresses multiple such regulators (Cardoso-Moreira et al. 2019). I will assume that a gene is expressed if its promoter harbors a high-affinity (“active”) binding site, i.e. a DNA sequence that can be recognized by at least one expressed transcriptional activator, thus equating active genes with active binding sites. I will
consider binding sites of a fixed length \( L \), and assume that the regulatory DNA that determines a gene’s expression is no longer than these \( L \) nucleotides, i.e. that a binding site is not embedded in a much longer region of potentially regulatory DNA. Any one binding site of length \( L \) exists in a sequence space of \( 4^L \) nucleotides. In any one organism, tissue, or cell, and in any one environment, some number \( B \) and proportion \( p_B = B/4^L \) of \( L \)-mers in sequence space are binding sites for TFs that are expressed.

Each nucleotide of a binding site can be affected by DNA mutations, which occur at a rate \( \mu \) per nucleotide and generation. I assume that mutation rates are small, such that any one binding site is affected by only one mutation at a time, which can destroy a binding site (with probability \( \mu^- \)) or create a new binding site (\( \mu^+ \)). Elementary population genetics prescribes an analytical upper bound on the incidence of misregulation, i.e. the extent of misregulation to be expected in the absence of selection acting against it (Supplementary File S2). This incidence depends on \( \mu^- \) and \( \mu^+ \), which depend in turn on the fraction \( p_B \) of sequence space filled with binding sites. Thus, \( p_B \) is a crucial quantity affecting the incidence of misregulation. I will estimate \( p_B \) from empirical data in the next section. To avoid modeling binding sites for thousands of genes individually, I will group an organism’s TF practical data in the next section. To avoid modeling binding sites, I next turn to experimental data from practitioners. To estimate how considering only expressed TFs would reduce effects of selection data from humans for the same seven tissues (Uhlen et al. 2003; Froula and Francino 2007; Racimo and Schraiber 2014; Qian and Kussell 2016). However, the simulations show that selection does not reduce the equilibrium proportion of incorrectly active genes \( f_{01} \) as a function of the strength of selection \( s_{01} \) against such genes, for various fractions \( p_B \) of genotype space filled with binding sites. The effects of selection scale with population size (Kimura 1983), and I explore a broad range of selection strengths that ranges from very weak \( (s_{01} = 0.01/N_e) \) to very strong \( (s_{01} = 100/N_e) \). Even though individual binding sites evolve neutrally at \( s_{01} < 1/N_e \), I note that selection against individuals that harbor many wrongly active binding sites may still affect evolutionary dynamics (Hahn et al. 2003; Froula and Francino 2007; Racimo and Schraiber 2014; Qian and Kussell 2016). However, the simulations show that selection does not reduce the equilibrium proportion of incorrectly active genes appreciably until \( s_{01} > 1/N_e \). The horizontal line (“empirical”) indicates an available empirical estimate of the strength of selection against wrongly active binding sites \( s_{01} \approx 0.1/N_e \) (Hahn et al. 2003), which is far below this value. For selection this weak, \( f_{01} \approx 0.062(p_B = 0.05) - 0.42(p_B = 0.45) \).
Figure 1 TF binding sites occupy a substantial proportion of sequence space. Each panel shows the fraction \( p_B \) of sequence space occupied by high affinity TF binding sites (vertical axis) as a function of a given number of TFs (horizontal axis). That is, the vertical axis shows the number of DNA sequences that are bound by at least one of these TFs, normalized by the size of sequence space for \( L = 8 \). All data are based on the (possibly small) subset of TFs with binding sites of length \( L \leq 8 \) for which PBM data are available. Thus, the resulting estimates of \( p_B \) are underestimates. Panel (A) is based on TFs encoded in one of seven genomes (color legend), regardless of where these TFs are expressed. Panel (B) is based on human TFs expressed in 7 tissues (color legend), according to RNA sequencing (rpkm) data from Cardoso-Moreira et al. (2019). Panel (C) is based on human TFs expressed in the same seven tissues as in (B) but using proteomic data from Uhlen et al. (2015).

Figure 2B shows analogous simulation data for the fraction of wrongly inactive genes \( f_{10} \). In contrast to \( f_{01} \), \( f_{10} \) declines as \( p_B \) increases, because mutations become less likely to deactivate a binding site, and more likely to convert one active binding site into another active binding site. However, in other ways \( f_{10} \) behaves similarly to \( f_{01} \), declining strongly with increasing selection strength. At empirically estimated values of selection strength (horizontal line, “empirical,” \( s_{10} \approx 10/N_c \); Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009), \( f_{10} \) ranges from 2.8% (\( p_B = 0.45 \)) to 31.6% (\( p_B = 0.05 \)) of genes that are incorrectly off.

From the equilibria for \( f_{01} \) and \( f_{10} \), it is straightforward to calculate the equilibrium excess of wrongly active genes \( \Delta_m \). Figure 2C shows \( \Delta_m \) for an intermediate value of \( p_B = 0.25 \) and under the assumption that half of all genes must be expressed for optimal adaptation (\( f^0 = C_{on}/G = 0.5 \)). At empirically observed selection strengths (horizontal line, “empirical,” \( s_{10} \approx 10/N_c \); Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009), \( f_{10} \) ranges from 2.8% (\( p_B = 0.45 \)) to 31.6% (\( p_B = 0.05 \)) of genes that are incorrectly off. Specifically, at \( p_B = 0.25 \), \( \Delta_m = 0.08 \), i.e. there is an 8% excess of wrongly active genes. In a genome of \( 10^6 \) genes, this number would correspond to 800 more wrongly active than wrongly inactive genes. This excess is a consequence of stronger selection against wrongly inactive genes. If selection were to act equally strongly against wrongly active and wrongly inactive genes (\( s_{01} = s_{10} \)), then \( \Delta_m < 0 \). The reason lies in the mutational dynamics of binding sites. As long as less than half of sequence space comprises active binding sites (\( p_B < 0.5 \)), mutation is more likely to destroy active binding sites than it is to create active binding sites. Tilting this balance toward wrongly active genes requires stronger selection against wrongly inactive genes.

With increasing \( p_B \) this balance is further shifted toward a prevalence of wrongly active genes and thus increasingly positive \( \Delta_m \). For example, at \( p_B = 0.45 \), \( \Delta_m = 0.19 \) (Supplementary Figure S3). Another quantity that influences \( \Delta_m \) is \( f^0 \), the fraction of genes that must be active for optimal adaptation, and for a trivial arithmetic reason: Everything else being equal, the more genes that must be active, the greater will be the proportion of all genes that are wrongly inactive, and the smaller will be \( \Delta_m \) (Supplementary Figure S4).

**Misregulation can be beneficial under realistic but not universal conditions**

So far, I have focused on \( \Delta_m \), the first of three quantities affecting the consequences of misregulation. The second quantity is the
fraction \( f_N \) of genes that should be expressed under optimal adaptation, for example because the sign of \( \frac{2f_N}{c_0} \) affects whether misregulation increases or decreases the fraction of correctly expressed genes (Equation 4). The fraction of genes expressed in various tissues and in environments to which an organism is well-adapted generally exceeds 0.5, and often substantially so, such that generally \( \frac{2f_N}{c_0} > 0 \) will hold (Supplementary File S1, Figure S1).

The third important quantity is the expression correlation \( c \) between gene expression states in different environments or states. A high correlation \( c \) can reduce the benefit or detriment of misregulation (e.g. Equation 3). RNA sequencing and microarray data from mouse, humans, and lower eukaryotes suggests that this correlation usually exceeds \( c = 0.5 \) (Supplemental Files S6 and S7, Figure S1, B and C).

With these quantities in mind, I will next use my simulation data to explore how misregulation might affect gene expression in a new cell state or environment. I first consider a strength of selection against misregulation consistent with experimental data \( (s_0 = \frac{1}{N_e}; s_1 = \frac{10}{N_e}) \), and an intermediate fraction \( p_B = 0.25 \) of sequence space being filled by TF binding sites. I also assume that an intermediate proportion \( f_O = 0.5 \) of genes is required to be optimally expressed in the old state. I will examine the effects of misregulation as a function of the fraction \( f_N \) of optimally expressed genes in the new state, and of \( c \), the correlation in gene expression between the old and the new state.

Figure 3A focuses on the first effect, the change in the total number of correctly active genes, \( D^m = f_0 + c \frac{(1-f_N)}{c_0} \). I note again that usually \( f_N ; c/c_2 > 0.5 \). For the entire parameter range considered here, \( D^m > 0 \), i.e. misregulation provides a benefit by

Figure 2 Substantial misregulation exists in mutation-selection-drift equilibrium. Logarithmically (base 10) transformed fraction \( f_0 \) of wrongly active genes among all \( G_{off} \) genes, as a function of the strength of selection \( s_{01} \) against such genes (horizontal axis), and of the fraction \( p_B \) of sequence space filled with TF binding sites (color legend). (B) Like (A), but for the fraction \( f_{10} \) of wrongly inactive genes among all \( G_{on} \) genes, as a function of the strength of selection \( s_{10} \) against such genes. (C) The excess of wrongly active genes, i.e. \( \Delta_m = f_0 - f_{10} \), as a function of the strength of selection against wrongly active genes (horizontal axis) and wrongly inactive genes (color legend). For panel (C), \( p_B = 0.25 \). Horizontal bars ("empirical") indicate the strengths of selection supported by empirical data (Hahn et al. 2003; Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009). All simulations are based on populations with \( N_e = 10^5 \) individuals, \( G = 1500 \) loci, \( G_{on} = 750 \), a mutation rate per nucleotide of \( \mu = 10^{-5} \), and an incidence of mutations leading to the destruction (\( \mu^\mu \)) or creation (\( \mu^+ \)) of binding sites estimated from mouse PBM data, as described in Materials and Methods. I initialized populations with zero misregulation \( (f_{01} = f_{10} = 0) \), and continued the simulations for \( 1/m \) generations, because preliminary simulations (not shown) had indicated that populations reach equilibrium by then. After \( 1/m \) generations, I calculated the population average of \( f_{01} \) and \( f_{10} \) over 100 generations. This average is the value shown in each bar chart. Error bars correspond to one standard deviation over these 100 generations, and are too small to be visible for most bars.
increasing the fraction of correctly active genes in a new state. The reason is that \( \Delta_0 > 0 \), i.e. more genes are wrongly active than wrongly inactive in the old state, because selection against wrongly active genes is weak. It is the wrongly active genes in the old state that help increase the fraction of correctly active genes in the new state. This benefit of misregulation increases with the total fraction \( f^f \), and for a simple reason: the more genes need to be expressed in the new state, the greater will be the positive contribution of genes that are active in the old state to \( \Delta_{\text{eff}} \). And the benefit also increases with the correlation \( c \) between the old and new states. As \( c \) increases, more of the genes that are expressed in the old state (wrongly or not) also need to be expressed in the new state. For the parameter range of Figure 3A, \( \Delta_{11} \) varies between zero and 0.08, that is, misregulation can increase the fraction of correctly expressed genes in a new environment by up to 8% of a genome’s genes.

The second effect of misregulation is the change in the total fraction of correctly inactive genes, \( \Delta_{00} = -\Delta_0(1-c)(1-f^s) \) (Figure 3B). It has the opposite sign of \( \Delta_{11} \). That is, as the excess \( \Delta_0 \) of wrongly active genes increases, the fraction of wrongly inactive genes among all genes that are misregulated in the old environment decreases. Because these wrongly inactive genes contribute to the correctly inactive genes in the new environment, \( \Delta_{00} \) decreases as well. As a consequence of \( \Delta_0 > 0 \) for the parameters I consider, \( \Delta_{00} < 0 \), that is, misregulation reduces the fraction of correctly inactive genes in the new environment. Everything else being equal, an increasing fraction \( f^s \) of genes that need to be expressed in the new environment reduces this detriment of misregulation, for the simple arithmetic reason that high \( f^s \) implies a low fraction of genes that should not be expressed.

An increasing correlation \( c \) between expression states also decreases this detriment. As \( c \) increases, so does the fraction of genes that should be off in the new environment among those genes \( 1-f^s \) that are off in the old environment. In consequence, as \( c \) increases, more and more of the genes that are already off in the old environment (wrongly or not), are among those genes that must be off in the new environment, which decreases the
impact of $\Delta_{m}$ on $\Delta_{m}^{sm}$. Overall, the figure indicates that misregulation can reduce the incidence of correctly inactive genes by up to 8%.

By adding the change in the fraction of correctly on and correctly off genes, I obtain the impact of misregulation on the total fraction of correctly expressed genes, $\Delta^{sm}_{m} + \Delta^{sm}_{m} = \Delta_{m}g_{m}(2f^{st} - 1(1 - c + c) + c$ (Figure 3C). This joint impact is smaller than $\Delta^{sm}_{m}$ and $\Delta^{sm}_{m}$, because misregulation affects these two quantities in opposite ways. The figure also shows that the correlation $c$ strengthens the effect of misregulation. The reason is easiest to see for the extreme case where $c = 1$, i.e. all the genes that are actually (not) expressed in the old state should (not) be expressed in the new state. In this case, no matter the amount of misregulation, all genes are correctly expressed in the new state, and thus the benefit of misregulation is maximal. I note that for realistic parameter values ($c > 0.5, f^{st} > 0.5$), the total fraction of correctly expressed genes is small but positive. Its maximal value is $\Delta^{sm}_{m} + \Delta^{sm}_{m} = 0.08$, i.e. misregulation can help increase the number of correctly expressed genes by at most 8%.

Finally, misregulation can affect fitness in the new environment (Equation 5). Because selection may act differently on wrongly off and wrongly on genes, the sign and magnitude of this effect cannot be directly inferred from the total fraction of correctly expressed genes. Figure 3D displays this fitness effect through the quantity $S_{m} = r_{m} - 1$, where $r_{m}$ is the ratio of fitness with and without misregulation from Equation (5). A positive value of $S_{m}$ indicates that misregulation increases fitness. The figure shows that misregulation increases fitness throughout the range of parameters shown. The reason is, again, that selection acts more strongly against wrongly off genes, effectively weighing the fraction of correctly on genes more strongly than that of correctly off genes. A value of $S_{m} = 0.5$ indicates that misregulation would increase fitness by 50% relative to the wild-type.

I notice that the large fitness effects in Figure 3D are partly due to computational feasibility restrictions, which allowed me to simulate only small populations ($N_{s} = 1000$). Because the strength of selection is scaled by $N_{s}$, smaller $N_{s}$ implies stronger selection in absolute terms. For example, a selection coefficient $s_{0} = 10/N_{s}$ has an absolute value of $s_{0} = 10^{-2}$ in a population with $N_{s} = 10^{7}$, but a value of $s_{0} = 10^{-4}$ in a population with $N_{s} = 10^{5}$. Because the fitness function, I use here is of the type $w = (1 - s)^{G}$, where $G$ is a number of loci, a larger absolute value of $s$ means a larger effect on fitness. The fitness effects of misregulation will thus be magnified in small populations. For example, everything else being equal in Equation (5), a 40% fitness increase at $N_{s} = 10^{5}$ would translate into an approximately 0.4% increase at $N_{s} = 10^{7}$. This more modest increases would remain visible to natural selection, because selection is more effective in large populations (Kimura 1983). For these reasons, I focus here on the sign of the fitness effect rather than on its magnitude.

The qualitative relationships, I describe above are sensitive to the parameters that influence $\Delta_{m}$ and its sign. As the fraction $p_{b}$ of sequence space filled with binding sites decreases, $\Delta_{m}$ becomes less positive, $\Delta^{sm}_{m}$ decreases, $\Delta^{sm}_{m}$ increases, and the total fraction of correctly expressed genes, together with the fitness benefit of misregulation decreases, until the sign of this benefit becomes negative at the smallest values of $p_{b} = 0.05$ I consider here (Supplementary Figure S5). Conversely, as the fraction $p_{b}$ of sequence space filled with binding sites increases, $\Delta_{m}$ becomes more positive, $\Delta^{sm}_{m}$ increases, $\Delta^{sm}_{m}$ decreases, and the total fraction of correctly expressed genes, together with the fitness benefit of misregulation increases (Supplementary Figure S6). Also, as the fraction $f^{st}$ of genes that need to be expressed for optimal expectation increases, $\Delta_{m}$ decreases in magnitude until it becomes negative, inverting all of the above relationships. Supplementary Figure S7 shows an example for $p_{b} = 0.25, f^{st} = 0.75$, where $\Delta_{m} < 0$, and where misregulation increases only the fraction of correctly off genes $\Delta^{sm}_{m}$. It decreases the fraction of correctly on genes, the total fraction of correctly expressed genes and mean fitness.

In sum, for a fraction of sequence space filled with binding sites and for selection strengths that are consistent with empirical data, misregulation can increase the fraction of correctly expressed genes and fitness in a new state. Whether it does so depends on the fraction of genes that need to be optimally expressed in a given state, and on the correlation between gene expression across states, which may differ among organisms and environments.

**Discussion**

I derived general mathematical expressions for how misregulation and crosstalk can increase or reduce adaptive gene expression in a new organismal state or environment. I then used these expressions, in combination with experimental data and computer simulations to ask how misregulation affects the total fraction of correctly active genes, correctly inactive genes, and correctly expressed genes, as well as a population’s mean fitness. When selection acts more strongly against wrongly inactive genes than against wrongly active genes (Hahn et al. 2003; Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009), misregulation can help increase the fraction of correctly expressed genes and mean fitness in a new environment or organisinal state, unless the fraction of sequence filled with binding sites is very small, and unless much more than half of all genes must be expressed for optimal adaptation.

A key quantity in determining whether misregulation is beneficial or detrimental is the difference between the fraction of incorrectly active and incorrectly inactive genes ($\Delta_{m} = f_{O}^{st} - f_{I}^{st}$). For example, $\Delta_{m}$ must be positive if misregulation is to help increase the total fraction of correctly active genes (Equation 2). The reason is that some of the wrongly active genes in an old state contribute to the correctly active genes in the new state, and the more such genes exist, the greater the chances that the net contribution of misregulation is positive. As long as less than half of sequence space is filled with active TF binding sites ($p_{b} < 0.5$), mutation pressure favors the destruction of binding sites (Supplementary File S2). Thus, when selection against misregulation is weak or absent, more genes will be wrongly inactive than wrongly active ($\Delta_{m} < 0$) in mutation equilibrium. This balance can shift to an excess of wrongly active genes ($\Delta_{m} > 0$) only when (i) $p_{b} > 0.5$, or (ii) when selection acts more strongly against wrongly inactive genes than against wrongly active genes ($s_{10} > s_{01}$), as has been empirically observed (Hahn et al. 2003; Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009).

When a minority of genes must be expressed for optimal adaption in a new state $f^{st} < 0.5$, some of these relationships can become inverted, i.e. an excess of wrongly inactive genes ($\Delta_{m} < 0$) can become favorable (Equation 4). I did not explore this scenario in detail, because most available expression data shows that organisms express more than half of their genes most of the time (Rasmussen et al. 2009; Haas et al. 2012; Wang et al. 2019).

A second key quantity is the fraction $p_{b}$ of sequence space filled with TF binding sites, because it affects $\Delta_{m}$. Under the conditions examined here, misregulation in an old environment or state is more likely to become adaptive as $p_{b}$ increases (Figure 3,
Supplementary Figures S5 and S6). In addition, \( p_B \) also affects the incidence of misregulation and crosstalk themselves. Crosstalk can be rare when TF binding sites are rare in sequence space. However, this is not the case. PBM data, in combination with gene expression data from multiple tissues and organisms in mouse and humans (Uhlen et al. 2015; Cardoso-Moreira et al. 2019) suggest that at least 5–30% of sequence space is filled with high affinity TF binding sites that are bound by at least one TF expressed in any one tissue (Figure 1). This may be a substantial underestimate, because the necessary data are available only for a subset of TFs. It is also based on TFs with short binding sites (\( L \leq 8 \) nucleotides). Although they constitute a majority of all TF binding sites (between 12% in humans and 40% in the round worm, see Materials and Methods), such short binding sites contribute disproportionately to misregulation, because they are especially likely to arise by chance alone through random mutations of genomic DNA. Notably, most sequences in sequence space are only one mutation away from a regulatory sequence (Supplementary Figure S2), and such "presites" can greatly accelerate the evolution of new binding sites for adaptive or nonadaptive reasons (MacArthur and Brookfield 2004; Tugrul et al. 2015). In addition, my simplifying assumption that a gene’s regulatory region is no longer than the length \( L \) of a TF binding site means that my estimates of the fraction of sequence space filled with regulatorily active DNA sequences are lower bounds (Supplementary Figure S8). More realistic assumptions about regulatory regions would broaden the conditions under which misregulation in an old environment is adaptive in a new one.

The dense filling of sequence space with regulatory signals is not restricted to higher eukaryotes. In yeast, 83% of random DNA fragments (\( L = 80 \) bp) can drive gene expression (de Boer et al. 2020). Regulatory signals are even abundant in prokaryotes, even though their transcriptional regulators have longer and information-rich binding sites (mean \( L \approx 16 \) bp, Stewart et al. 2012). For example, 7% of random synthetic DNA fragments (\( L \approx 150 \) bp) can drive the expression of a reporter gene in E. coli (Urtcho et al. 2020). Experiments that replaced the promoter of the lac operon with random synthetic DNA fragments (\( L \approx 100 \) bp) showed that 10% of such sequences could drive the operon’s expression in response to lactose. In the same experiments, an additional 60% of sequences contained pre-sites, i.e. they were only one nucleotide change away from a regulatory sequence (Yona et al. 2018). More generally, experimental evolution can easily create prokaryotic regulatory DNA from random DNA fragments (Hornitz and Loeb 1986; Wolf et al. 2015; Yona et al. 2018; Urtcho et al. 2020). Taken together, this evidence suggests that \( p_B \) is not even small in prokaryotes, and that their regulatory signals are generally not sparse in sequence space. If so, misregulation and crosstalk will be frequent in many organisms, rendering their detriments and benefits an important subject of study. Estimating \( p_B \) for prokaryotes quantitatively remains an exciting task for future work.

One major limitation of this work is the use of simulations to study the evolutionary dynamics of misregulation, because the little pertinent analytical theory (Woodcock and Higgs 1996; Stewart et al. 2012) can give misleading results when applied to thousands of loci (Woodcock and Higgs 1996). The development of adequate theory is an important task for future work. Such theory is also necessary to overcome some of the simplifying assumptions I made here. One of them is that I considered only transcriptional activation but not repression (Grah and Friedlander 2020). Another is that individual TF binding sites contribute independently to fitness, whereas their effects are often interdependent. For example, multiple mutations often affect fitness more strongly (synergistic epistasis) or more weakly (antagonistic epistasis) than expected from their individual effects (Kouyos et al. 2007), which can either increase or decrease the efficacy of selection against misregulation. Another simplifying assumption regards the absence of recombination, which could increase the efficacy of selection against misregulation by bringing multiple mutant TF binding sites causing gene misexpression together in the same genome. A fourth assumption is that genes can only be on or off, whereas they are actually expressed at different levels that span several orders of magnitude in mRNA or protein concentrations (Haas et al. 2012; Cardoso-Moreira et al. 2019; Wang et al. 2019). Relatedly, I assumed that TF binding to DNA is all-or-nothing, whereas actual binding is well described by thermodynamic models as a function of TF concentration and affinity for a specific binding site (Berg et al. 2004; Bintu et al. 2005a, 2005b; Mustonen and Lassig 2005; Friedlander et al. 2016). For continuously expressed genes, misregulation implies that individual genes are expressed at a level that is higher or lower than optimal. Modeling continuous expression genome-wide is especially challenging, because a gene’s expression level is not equivalent to the strength of selection acting on it. To be sure, highly expressed genes are often important, as inferred from their low tolerance for mutations on laboratory or evolutionary time scales. However, this association is not strong, and lowly expressed genes can also be subject to strong selection (Pal et al. 2003; Wall et al. 2005; Vitkup et al. 2006; Lehner 2008).

In addition to better theory, improved predictions will also need more data. For example, PBM and gene expression data for more TFs will help improve estimates of the fraction \( p_B \) of sequence space filled by binding sites for expressed TFs. In addition, current estimates of the strength of selection against misregulation are limited (Hahn et al. 2003; Mustonen and Lassig 2005; Mustonen et al. 2008; Kim et al. 2009). They generally estimate selection coefficients as genome-wide averages, whereas the strength of selection may vary widely among genes (Racimo and Schraiber 2014). Despite the importance of such data, I note that theoretical work will remain essential to study the evolution of misregulation and its consequences in the foreseeable future, because of limitations in experimental technology. For example, selection against individual misexpressed genes is sufficiently weak (\( s < 0.1/N_e\) or \( 10^{-6}\)), that current technologies cannot detect it for organisms with typical population sizes of \( N_e > 10^6\), although we know that it is strong enough to leave genomic signatures on evolutionary time scales (Hahn et al. 2003; Froula and Francino 2007; Racimo and Schraiber 2014; Qian and Kussell 2016). In other words, experiments cannot tell us whether any one gene is optimally expressed, with the exception of few genes whose misexpression has dire consequences, where experiments can manipulate misregulation by mutating regulatory regions and measuring the ensuing fitness effects.

A further limitation of this work is the tacit assumption that a single high-affinity binding TF suffices to express a nearby gene. Such a site is necessary but not sufficient for gene activation, which also requires DNA accessibility (open chromatin), correct DNA conformation (Mathelier et al. 2016), as well as the right methylation status (Bird 1995). Other important factors include a site’s distance to the transcription start site, its orientation on the DNA helix, and the presence of essential co-activators, which may or may not themselves bind DNA (Alberts et al. 2008; Long et al. 2016). Perhaps the most important complication is cooperative gene regulation by multiple binding sites for the same factor, and combinatorial regulation by binding sites for different factors.
Such complex regulation could in principle completely prevent misregulation. However, several lines of evidence suggest that this is not the case. First, theory suggests that combinatorial regulation only reduces the extent of misregulation and may not even do so dramatically (Friedlander et al. 2017; Grah and Friedlander 2020). For example, combinatorial regulation can reduce misregulation substantially when it results in fewer TFs needed by an organism, but the number of TFs increases faster than the total number of genes in a genome (van Nimwegen 2006), showing that combinatorial regulation does not operate close to optimality. Second, regulatory regions, especially in higher eukaryotes can extend over multiple kilo base pairs (kbp), which can greatly increase chance occurrences of multiple TF binding sites and thus spurious gene expression. For example, even with $p_0$ as low as 0.05 ($L = 8$), a single random 1 kbp stretch of DNA is expected to contain 28 high-affinity TF binding sites (Supplementary Figure S8). Third, the more binding sites are involved in a gene’s regulation, the less specific and more degenerate individual sites become (Stewart and Plotkin 2013). Fourth, complex eukaryotic gene regulation is often mediated by low affinity binding sites (Tanay 2006; Jaeger et al. 2010; Crocker et al. 2015, Crocker and Preger-Ben Noon 2016), which exist in vastly greater numbers than the high affinity sites I consider, and are much more likely to occur by chance alone in noncoding DNA. In sum, rather than eliminating misregulation altogether, combinatorial regulation may serve to keep it in check as genomes become larger, as they encode more TFs, and as sequence space becomes increasingly filled with TF binding sites. Models involving combinatorial regulation and additional factors require many additional assumptions, for example about gene regulatory logic (Buchler et al. 2003; Wagner et al. 2007), and remain an important challenge for future work.

Although my model focuses on genes, it could also be applied to any transcribed genomic DNA. In this form, it could help answer the question whether regulatory crosstalk can facilitate the de novo origin of genes, which often involves new transcription (Begun et al. 2007; Zhao et al. 2014; Ruiz-Orrera et al. 2015; Neme and Tautz 2016). In higher eukaryotes, most genomic DNA is transcribed (Djebal et al. 2012), but only a minority of the genome is subject to selection (<10% in mammals, Lindblad-Toh et al. 2011, Graur et al. 2013). This means that the fraction of transcripts needed for optimal adaptation may be small ($\theta^0, \theta^1 < 0.5$), that many spurious transcripts exist, and that selection against them is very weak, which suggests a (potentially large) excess of incorrectly transcribed genomic regions ($\Lambda_m > 0$). Under these conditions Equation (4) implies that misregulation reduces the fraction of correctly expressed transcripts in a new organismal state or environment, although it may increase fitness if some spurious transcripts provide a strong enough benefit to be preserved by natural selection.

An elementary question about any adaptive trait is whether the trait could invade a population when rare, e.g. when introduced in the form of a single mutant individual. For misregulation on a genome-wide scale, however, this question would be misguided. Such misregulation not only arises through a balance of selection and mutations at thousands of regulatory regions, it also reduces fitness in an environment or state to which an organism is well-adapted. In other words, misregulation has probably not evolved to increase fitness or to facilitate the adaptive evolution of gene expression. Whenever misregulation may be adaptive, this adaptiveness is a by-product of its existence. In this respect, misregulation is in good company with other phenomena that facilitate evolution. Consider enzyme promiscuity, in which an enzyme catalyzes one main reaction subject to natural selection but also one or more side reactions (O’Brien and Herschlag 1999; Khersonsky and Tawfik 2010). An enzyme’s efficiency at catalyzing the main reaction may trade-off with the side reactions (Bershtein and Tawfik 2008; Khersonsky and Tawfik 2010), such that very strong selection for this efficiency can lead to the elimination of the side reactions during evolution. Conversely, when selection on the main reaction is weak, the side reactions can persist and facilitate survival in a new environment, as has been shown for enzymes that include antibiotic resistance proteins (Aharoni et al. 2005; Bershtein and Tawfik 2008). A similar principle holds for a main driver of Darwinian evolution, DNA mutation itself. Because most mutations are deleterious, mutations will reduce a population’s mean fitness, and natural selection will favor a small mutation rate unless an organism experiences a new environment in which many mutations may be beneficial. However, selection is not sufficiently strong to reduce the mutation rate to zero, and especially so in complex organisms with small populations where selection is less effective than in larger populations. Such organisms may also experience higher mutation rates (Lynch et al. 2016) and with them, a greater incidence of the occasionally beneficial mutation.

Misregulation is an analogous phenomenon affecting gene expression, which we know is involved in many adaptations (Gasch et al. 2000; Carroll et al. 2001). Like DNA mutations and enzyme promiscuity, misregulation is one of life’s many imperfections that can help propel Darwinian evolution.

I would like to thank Simon Aeschbacher, Tamar Friedlander, Joshua Payne, Gabriel Schweizer, and Caua Westmann for valuable comments on an early draft of this study.

I acknowledge financial support from the European Research Council under Grant Agreement No. 739874, and from the Swiss National Science Foundation grant 31003A_172887, as well as from the University of Zurich Priority Research Program in Evolutionary Biology.

None declared.


Communicating editor: J. Masel