Sexual antagonism drives the displacement of polymorphism across gene regulatory cascades – Supporting Information

Mark S. Hill[1,2], Max Reuter[*2] & Alexander J. Stewart[**3]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

2 Research Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[3] Department of Biology & Biochemistry, University of Houston, Houston, TX, USA

[*] Email: m.reuter@ucl.ac.uk

[**] Email: astewar6@central.uh.edu

DOI: 10.1098/rspb.2019.0660

In this supplement we give further details of the modeling framework, simulations and analytical results used to produce the results presented in the main text. We begin by describing evolution of a binding site under a biophysical model for transcription regulation in the presence of sexually antagonistic selection on a target gene. We then describe the evolutionary dynamics of the model under weak mutation and the conditions that lead to polymorphism between pairs of alleles. These models and analytical approximations are derived for a single (diploid) target gene. Finally, we describe an extension to our model to include multi-gene cascades and the conditions that lead to displaced polymorphism in individual-based simulations under the diploid Moran process.

Gene regulation under weak mutation

We employ a simplified model of transcription regulation in which transcription factor binding sites are composed of a contiguous region of n nucleotides, such that at each position there is a "correct" nucleotide that contributes an amount ϵ to the site's binding energy, while an "incorrect" nucleotide contributes nothing. We assume that ϵ is constant across positions, and each position contributes equally and independently to the site's binding energy. The probability that the site is bound is then given by Eq. 1 in the main text. While in reality the assumption that each nucleotide position contributes equally and independently to binding energy does not necessarily hold [1, 2], this simplified model adequately describes the biophysics of transcription factor binding and generates a probability of binding that is (including the sigmoidal relationship between binding probability and the number of correct nucleotide matches) and therefore captures the evolutionary dynamics of gene regulation [1, 3, 4, 5, 6, 7].

The degeneracy among genotypes assumed by our simplified model allows us to reduce the number of different alleles associated with a given site from 4^n—the number of distinct genotypes that can occur—to $n+1$—the number of possible "correctly matched" nucleotides at the site. A binding site is thus characterized by a single number k which is the number of correctly matched nucleotides. Mutations occur through single nucleotide substitutions which increase, μ_k^+, or decrease, μ_k^-, the number of matches at the site from a starting value of k. The rates of these mutations are

$$\begin{aligned} \mu_k^+ &= \mu\frac{n-k}{3n} \\ \mu_k^- &= \mu\frac{k}{n} \end{aligned} \tag{S1}$$

where μ is the per-nucleotide rate of substitutions. What is immediately clear from Equation S1 above is that the rates of mutation increasing and decreasing the number of matches are asymmetrical and vary with genotype. This is a violation of the assumptions made in most simple population genetic models (of SA and more generally) and precludes the standard analytical treatment of allelic dynamics using a diffusion approximation (because the resulting system of differential equations cannot be solved). However, because we are considering mutations at the level of nucleotide substitutions with rates that are typically as low as $10^{-9} - 10^{-7}$ [8], we can treat the evolution of a binding site in the weak mutation limit, i.e., in the limit where the product of effective population size and mutation rate is small, or more accurately $2n\mu N_e \ll 1$.

In the absence of sexual antagonism, evolutionary dynamics in the weak mutation limit are well approximated if we assume that the population is monomorphic, and that a given mutation has time to either reach fixation or be lost before another arises. Thus, we can simply calculate the fixation probabilities of single mutations while ignoring the effects of clonal interference. If we write π_k^t for the probability that the population has a binding site with k matched nucleotides at time t, then in the weak mutation limit we can write

$$\pi_k^{t+1} = \pi_k^t(1 - \mu_k^+ \rho_{k\to k+1} - \mu_k^- \rho_{k\to k-1}) + \pi_{k+1}^t \mu_{k+1}^- \rho_{k+1\to k} + \pi_{k-1}^t \mu_{k-1}^+ \rho_{k-1\to k} \tag{S2}$$

where $\rho_{i\to j}$ is the probability that a mutant with j matches fixes in a population where a binding site with i matches is resident. For a given pair of alleles in the absence of sexual antagonism, this fixation probability is given by Kimura's expression [9]. At equilibrium, we can then use detailed balance to find the following recursion relation

$$\pi_k \mu_k^- \rho_{k\to k-1} = \pi_{k-1} \mu_{k-1}^+ \rho_{k-1\to k} \tag{S3}$$

which can be solved numerically.

When we are dealing with SA, this weak mutation treatment unfortunately breaks down. With balancing selection potentially acting on polymorphisms, allelic dynamics are too slow for populations to be assumed to be monomorphic and the evolutionary dynamics of an invading allele among males and females is in general different. We can get around this by making the additional assumption that selection is weak. Polymorphism then cannot typically be maintained for prolonged periods of time and an allele under antagonistic selection is at approximately equal frequencies in males and females [10]. Given this, we can use Equation S3 above to gain insight into the evolutionary dynamics of SA at a binding site under weak selection.

Mutation-selection gradient

To analyze the evolutionary dynamics of a binding site under in the weak mutation weak selection limit, first recall Kimura's [9] expression for the fixation probability of a mutation with relative fitness $1 + s$ against a resident with fitness 1:

$$\rho = \frac{1 - \exp[-2s]}{1 - \exp[-4Ns]} \tag{S4}$$

where weak selection requires $2Ns \ll 1$. We then use Equation 2 of the main text to calculate the average fitness effect of a mutation that changes the number of nucleotide matches in a binding site by ± 1:

$$s = \frac{w_m(k) + w_f(k)}{w_m(k+1) + w_f(k+1)} - 1 \tag{S5}$$

where k is the number of nucleotide matches in Equation 1 and $w(k)$ is the fitness for a homozygote with k nucleotide matches. Substituting this into Equation S3 we can calculate the ratio of transition probabilities for mutations that increase or decrease nucleotide matches

$$\phi_k = \frac{(n-k)\rho_{k \to k+1}}{3(k+1)\rho_{k+1 \to k}} \tag{S6}$$

If $\phi_k > 1$, the number of nucleotide matches tends to increase whereas when $\phi_k < 1$, the number of nucleotide matches tends to decrease. This can be used to describe the direction of evolutionary change of a binding site under uni-directional selection, as shown in Figure 3 of the main text.

Polymorphism

In order to measure polymorphism in a multi-allele system with a non-linear fitness we calculate the expression difference across alleles at each locus in each individual, which gives us the following expression for polymorphism p at a given locus:

$$p = \frac{2}{N} \sum_{i=1}^{N} |E_1^i - E_2^i| \tag{S7}$$

where E_j^i is the expression level of allele j (1 or 2) in individual i. If there are two alleles segregating at equal frequency, one with maximum expression and one with minimum expression, this will result in polymorphism $p = 1$ (since approximately half the population will be heterozygous and hence have $|E_1^i - E_2^i| = 1$). Note that this measure is chosen to be conservative, since high levels of polymorphism will only be encountered when there is both a high level of heterozygosity and a high level of expression difference between the alleles that are segregating. In contrast, we do not record populations as highly polymorphic if populations segregate for alleles that cause similar expression and hence negligible antagonistic fitness variation.

We used the results of [10] to determine whether a given pair of neighboring binding site variants segregating in a populations are favored to be polymorphic. This condition is given in [10] as

$$\frac{S_m(1-H_m)}{H_f(1-S_m)} > S_f > \frac{S_m H_m}{1-H_f+H_m S_m} \tag{S8}$$

where we have

$$\begin{aligned} S_m &= \frac{w_m(k+1,k+1)-w_m(k,k)}{w_m(k+1,k+1)} \\ S_f &= \frac{w_f(k,k)-w_f(k+1,k+1)}{w_f(k,k)} \\ H_m &= \frac{w_m(k,k)-w_m(k,k+1)}{w_m(k,k)-w_m(k+1,k+1)} \\ H_f &= \frac{w_f(k,k)-w_f(k,k+1)}{w_f(k,k)-w_f(k+1,k+1)} \end{aligned} \tag{S9}$$

We use Equations S8 and S9 together to determine whether polymorphism will arise in binding sites, as shown in Figure 3. We note that under both positive and negative curvature, dominance effects are in the same direction for both males and females, i.e. if $H_m > 0.5$, $H_f > 0.5$ and vice versa. To see this note that $H_m > 0.5$ implies

$$w_m(k,k+1) > \frac{w_m(k,k)+w_m(k+1,k+1)}{2}$$

In contrast to the polymorphic case, the conditions for fixation of a mutant allele in a single locus, two allele system are

$$S_f < \frac{S_m H_m}{1 - H_f + H_m S_m} \tag{S10}$$

for a male-beneficial mutation and

$$S_f > \frac{S_m(1 - H_m)}{H_f(1 - S_m)} \tag{S11}$$

for a female-beneficial allele. This is used to describe the evolutionary dynamics for a male- or female-beneficial binding site variant as shown in Figure 3 of the main text.

Continuum limit

To understand the evolutionary dynamics of binding sites described in Figure 3 of the main text it is instructive to consider the limit of continuous gene expression and small mutations. In this case Equation 9 above becomes

$$\begin{aligned} S_m &= \frac{\frac{dw_m}{dk}}{w_m(k,k)} \\ S_f &= -\frac{\frac{dw_f}{dk}}{w_f(k,k)} \\ H_m &= \frac{1}{2} \\ H_f &= \frac{1}{2} \end{aligned} \tag{S12}$$

and the condition for invasion of a male beneficial allele becomes

$$w_m \frac{dw_f}{dk} + \frac{dw_f}{dk}\frac{dw_m}{dk} < w_f \frac{dw_m}{dk} \tag{S13}$$

If we then take Equation 2 of the main text under the limit of strong selection, symmetrical selection, i.e. $h_m = h_f$, $s_m = s_f = 1$ we recover $\frac{dw_f}{dk} = -h_f w_f(1 - w_f)$ and $\frac{dw_m}{dk} = h_m w_m(1 - w_m)$ to give

$$(1 - w_f) < w_f(1 - w_m) \tag{S14}$$

for fixation of male-beneficial mutations and

$$w_m(1 - w_f) > (1 - w_m) \tag{S15}$$

for fixation of female-beneficial alleles. Figure S1 shows the resulting evolutionary dynamics that lead to polymorphism under fitness landscapes with negative curvature. Figure S2 shows an example trajectory for a population evolving under negative curvature. We see that evolutionary trajectories tend towards intermediate expression before gaining polymorphism, just as in Figure 3 of the main text. In the case of negative curvature this results in a decline in population mean fitness as male and female fitness equalize (Fig. S2).

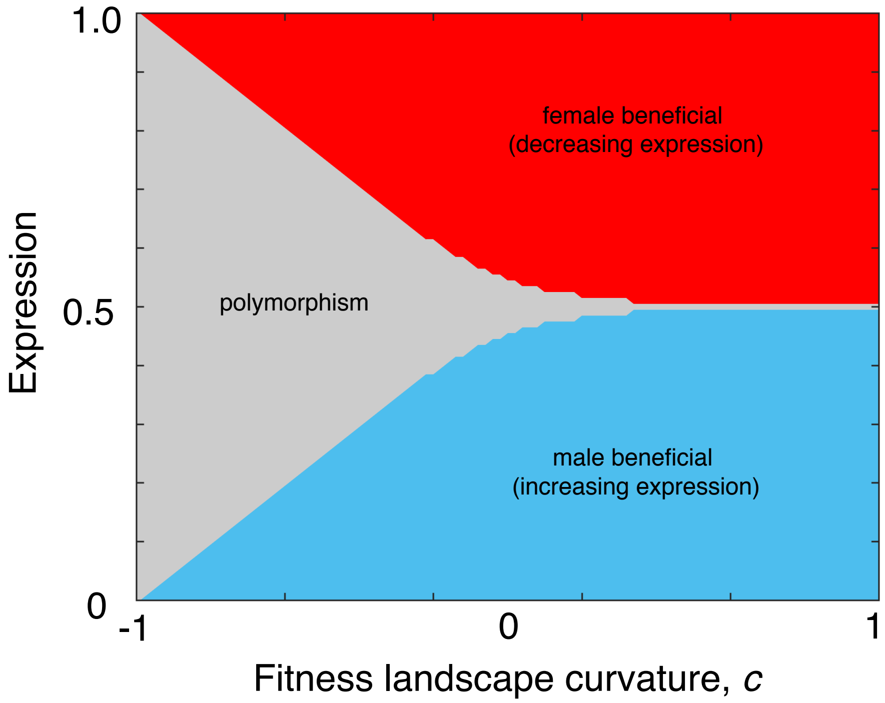

Figure S1: Evolutionary dynamics in the continuum limit. We determined the selection gradient on gene expression using Equations S14 and S15 for the continuum limit under strong selection. We see that expression tends towards intermediate values regardless of landscape curvature, with a large region of disruptive selection leading to polymorphism when curvature is negative.

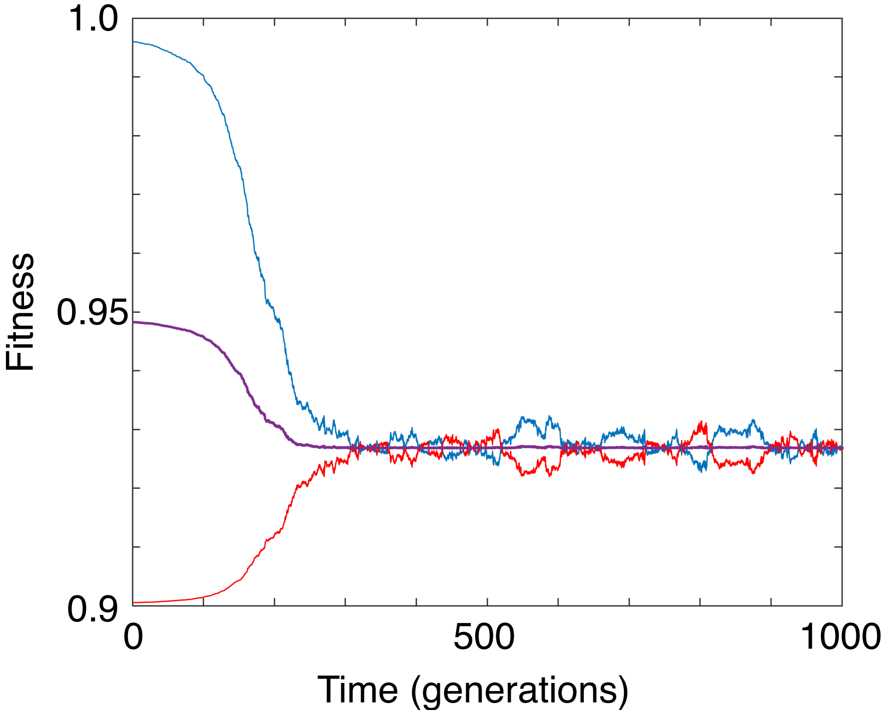

Figure S2: Sample path in the continuum limit. We ran an individual-based simulation in the continuum limit with a population size of N=1000, $s_m = s_f = 0.1$, $h_m = h_f = 10$ and $c_m = 1 - c - f = 0.6$. We see that mean fitness (purple) declines as male (blue) and female (red) fitness equalize.

Individual-based simulations of regulatory evolution

We carried out evolutionary simulations for a single gene and multi-gene cascades using the model for transcription factor binding described in the main text. We simulated populations evolving under the diploid Moran model [11] with sexual reproduction. Populations are composed of $N/2$ diploid males and $N/2$ diploid females. The model captures the case of a fixed population size with non-overlapping generations.

The model consists of birth-death events at each time step. Reproduction events occur by choosing one male and one female according to their normalised fitness. These two individuals produce a single offspring who receives one allele from each parent at each locus. The offspring is randomly assigned male or female sex with equal probability and another individual of the same sex is randomly selected (with uniform probability) to die.

Mutation events occur during the transmission of alleles from parents to offspring, with a per nucleotide mutation rate of $\mu = 0.1/(2Nn)$ to ensure weak mutation. We assume that no recombination events occur within TF binding sites, which is justified owing to the short sequence lengths under consideration [6]. In the case of simulated gene cascades, however, recombination can occur between the binding sites of the different genes.

We simulated regulatory evolution under SA selection on the expression of a gene. For cascades, the gene under selection is the terminal gene. We calculated the expression level of each gene in the cascade according to our model of TF binding which then gave the fitness of the individual based on the expression level of the terminal gene. Figure S3 shows the ensemble mean fitness over time for a three-gene cascade as described in the main text. We see dynamics similar to Figure S2 with male and female fitness equalizing at the expense of declining population mean fitness. However we the see a subsequent increase in mean fitness as displacement occurs (see Figure 4 of the main text).

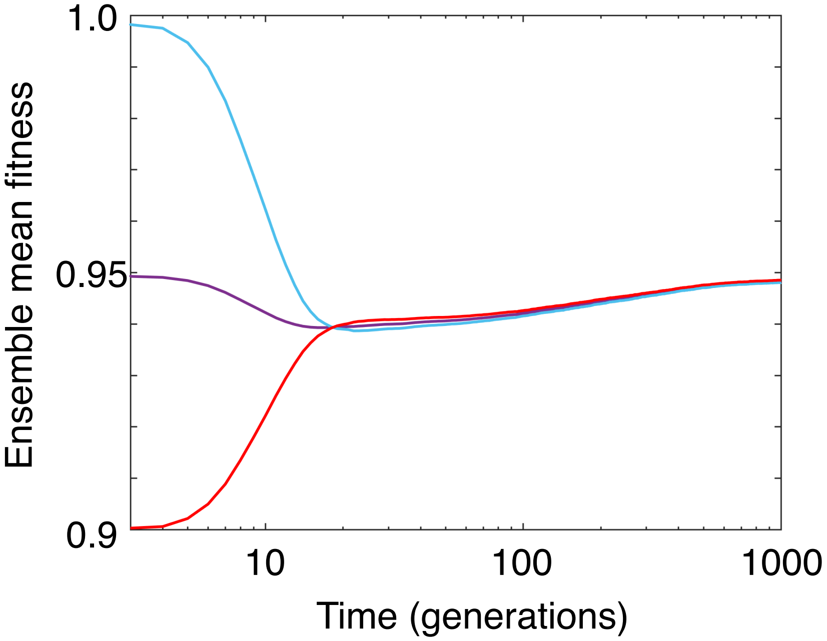

Figure S3: Ensemble mean fitness in a cascade. We plotted the ensemble mean fitness fitness for the whole population (purple), males (blue) and females (red) for the regulatory cascade described in Figure 4 of the main text. We see similar dynamics to Figure S3, with mean fitness declining as male and female fitness equalize. However we also see a subsequent increase in mean fitness and further convergence of male and female fitness as displacement starts to occur.

Displaced polymorphism

In the evolutionary dynamics described in Figure 4 of the main text we see polymorphism arise first at the terminal gene under direct SA selection, before being displaced to regulatory genes higher up the cascade. In particular, there is a tendency for polymorphism to be displaced ultimately to the highest gene in the cascade. This displacement of polymorphism raises three questions. First, why does polymorphism tend to arise at the terminal gene initially? Second, why does polymorphism get displaced? And third, why does polymorphism ultimately rise to the top of the cascade?

The initial emergence of polymorphism at the terminal gene can be explained by the effect of mutations at the three levels of the cascade on terminal expression levels. We assume initially that all binding sites are functional and genes are highly expressed. As shown in Figure 3 of the main text and Figure S1 above, selection is then essentially directional and favors decreasing expression at the terminal gene. This is most effectively achieved by decreasing binding strength at the terminal gene, since mutations at points higher in the cascade are of smaller effect. Regulatory variants therefore invade most likely at the binding site of the terminal gene and disruptive selection then leads to polymorphism at the bottom of the cascade.

Once regulatory polymorphism is established at the terminal gene, displacement occurs to mitigate the deleterious effects of intermediate expression in heterozygotes. When polymorphism resides at the terminal gene, heterozygotes with one strong and one weak terminal binding site have expression of approximately $E = 1/2$. In the fitness landscape with negative curvature, this is the lowest possible fitness. If polymorphism resides at a point higher in the cascade, in contrast, it is either gene 2 or gene 3 that has expression $E = 1/2$ in the heterozygote. This typically results in expression other than 1/2 in the terminal gene, and therefore greater heterozygote fitness. It is this fitness difference that drives the displacement of polymorphism. Accordingly, there is an increase in mean fitness as polymorphism gets displaced (see Fig. S3 and Fig. 4 of the main text) and cascades with polymorphism residing further up have higher fitness (Figs S4 and S5). Obviously, the advantage of superior heterozygote fitness would also be expected to favour the initial establishment of polymorphism at gene 2 or 3, rather than the terminal gene. But as explained above, this is less likely to occur due to the smaller mutational effects, and hence reduced probabilities of invasion, of regulatory mutations in the bindings sites of genes 2 and 3.

A remarkable result of our simulations is the fact that polymorphism always tends to rise to the top of the cascade and ultimately reside at gene 3 (Fig. 4, main text). At first sight, this is surprising, because there is no mean fitness advantage to polymorphism at gene 3 as compared to gene 2 (see Fig. S4). The difference in polymorphisms at gene 2 and gene 3 is related to the resulting sex-specific fitness, as shown in Figures S5 and S6. Here we can see that polymorphism residing at gene 2 typically results in a greater difference between male and female fitness than polymorphism resulting at gene 3. This implies that polymorphism at gene 2 does not totally remove sexual conflict, and as a result there is still disruptive selection that can favour displacement of polymorphism higher up the cascade. Polymorphism at gene 3 then allows for further balancing of male and female fitness, because there is more scope for "fine tuning" expression via gene 2.

The resulting dynamics of polymorphism in a cascade are summarized in Figure S7—there is an inevitable displacement away from the terminal gene as this results in greater heterozygote fitness. There is also a tendency to displace polymorphism to gene 3 rather than gene 2 as this reduces overall sexual conflict via fine tuning of expression lower in the chain.

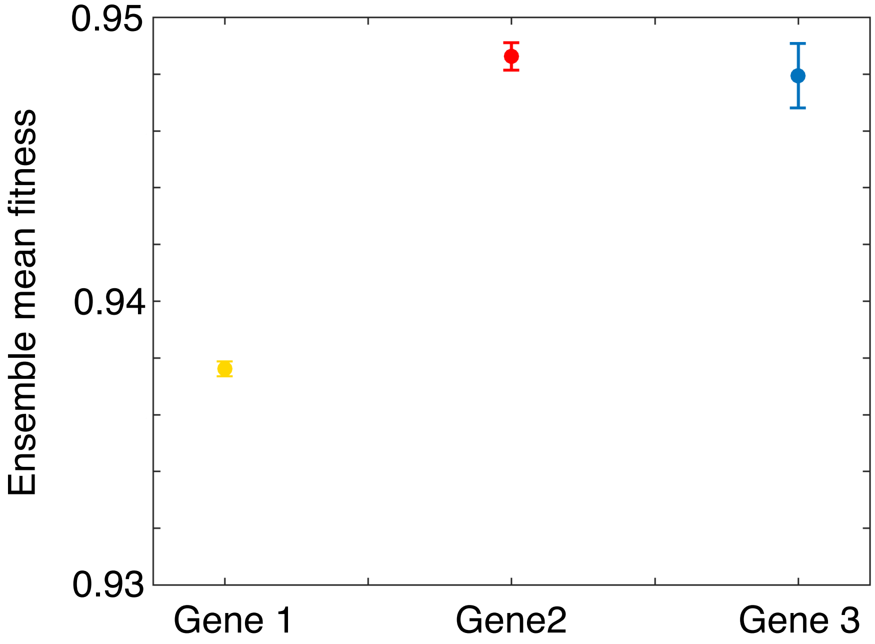

Figure S4: Mean fitness with polymorphisms at different positions in the cascade. We calculated the mean fitness for our cascade simulations for cases where polymorphism resides at the terminal gene, gene 2 or gene 3. We see a clear fitness advantage to polymorphism at genes 2 and 3 compared to the terminal gene 1, but no advantage of polymorphism at gene 3 over polymorphism at gene 2.

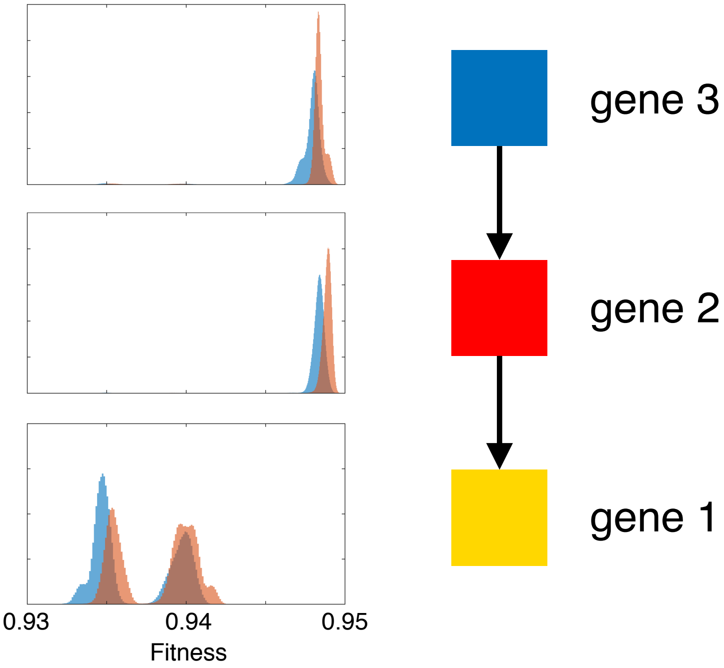

Figure S5: Male and female fitness with polymorphisms at different positions in the cascade. We plotted the distribution of male (blue) and female (red) fitness for populations with polymorphism residing at gene 1, gene 2 and gene 3 respectively. We see that gene 1 has both lower average fitness and larger fitness differences between the sexes. Polymorphism at gene 2 results in higher mean fitness and smaller but still prevalent fitness differences between the sexes. Polymorphism at gene 3 results in the lowest fitness difference between the sexes.

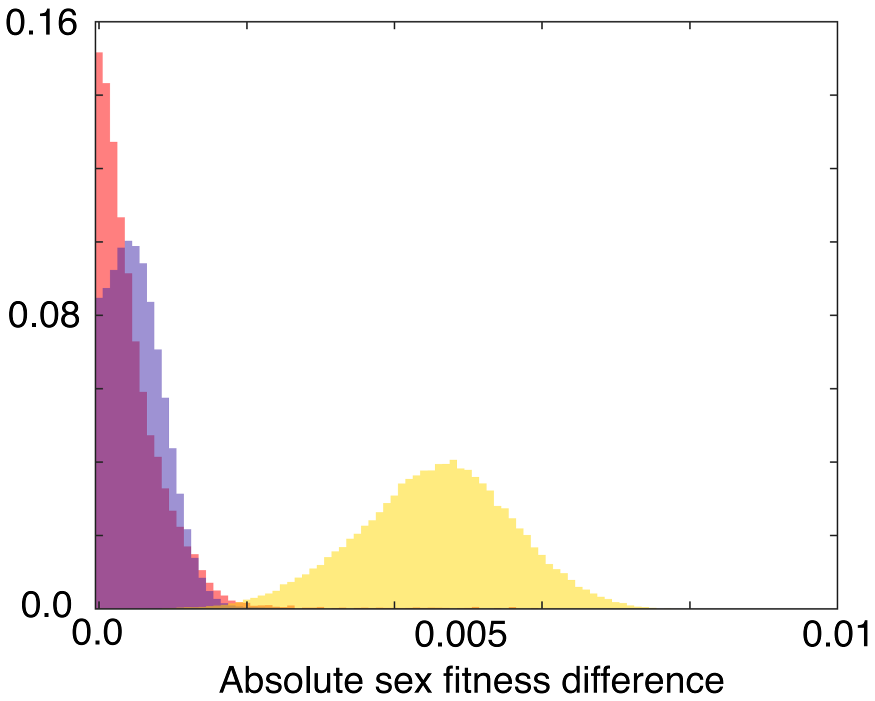

Figure S6: Difference between male and female fitness with polymorphisms at different positions in the cascade. We plotted the absolute fitness difference between males and females for each population with polymorphism residing at gene 1 (yellow), gene 2 (blue) or gene 3 (red). We see that gene 1 always maintains large differences, while gene 2 eliminates differences in some but not most cases, whereas gene 3 the smallest fitness difference between the sexes.

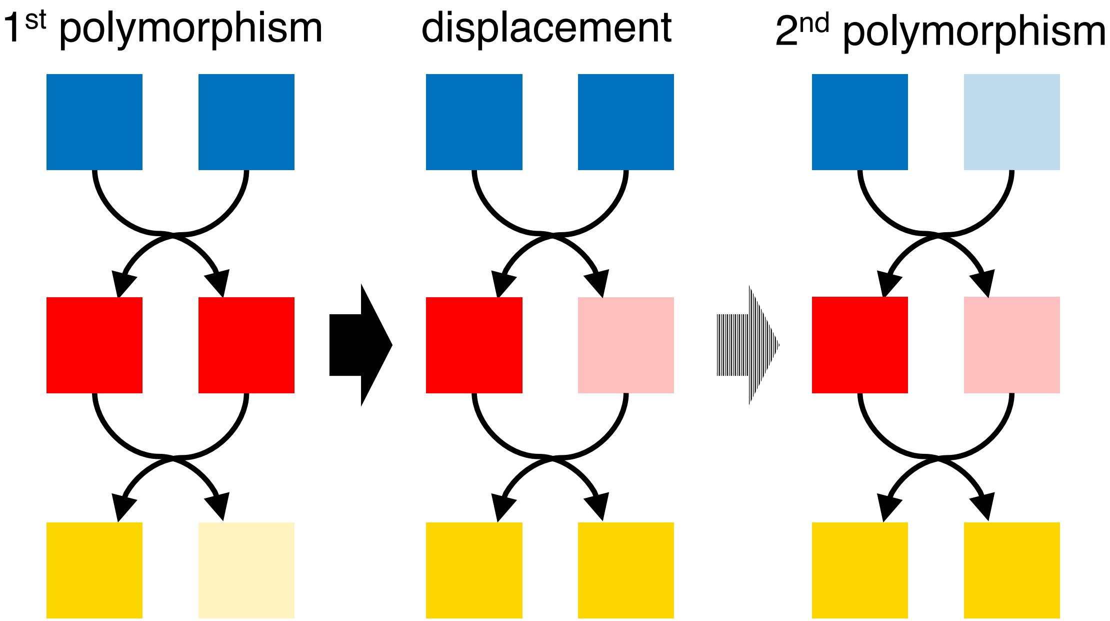

Figure S7: The dynamics of displaced polymorphism. Our study reveals a typical sequence of evolutionary responses to sexually antagonistic selection on gene regulation are as follows. (Left) Polymorphism arises first at the gene under SA selection, where selection is strongest. (Center) polymorphism subsequently gets displaced up the cascade as this delivers a fitness benefit and tends to reduce conflict by equalizing male and female fitness. (Right) When polymorphism is displaced to intermediate levels in the cascade, a second displacement is likely to occur to a point higher in the cascade, as this tends to further reduce conflict.

References

[1] Gerland, U. & Hwa, T., 2002 On the selection and evolution of regulatory DNA motifs. *Journal of Molecular Evolution* **55**, 386–400. ISSN 00222844. (doi:10.1007/s00239-002-2335-z).

[2] Lässig, M., 2007 From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* **8**, S7. ISSN 14712105. (doi:10.1186/1471-2105-8-S6-S7).

[3] Buchler, N. E., Gerland, U. & Hwa, T., 2003 On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5136–41. ISSN 0027-8424. (doi:10.1073/pnas.0930314100).

[4] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J. & Phillips, R., 2005 Transcriptional regulation by the numbers: Models **15**, 116–124. ISSN 0959437X. (doi:10.1016/j.gde.2005.02.007).

[5] Mustonen, V., Kinney, J., Callan, C. G. & Lässig, M., 2008 Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 12376–12381. ISSN 1091-6490. (doi:10.1073/pnas.0805909105).

[6] Stewart, A. J., Hannenhalli, S. & Plotkin, J. B., 2012 Why transcription factor binding sites are ten nucleotides long. *Genetics* **192**, 973–85. (doi:10.1534/genetics.112.143370).

[7] Tuğrul, M., Paixão, T., Barton, N. H. & Tkačik, G., 2015 Dynamics of transcription factor binding site evolution. *PLoS Genet* **11**, e1005639. (doi:10.1371/journal.pgen.1005639).

[8] Nachman, M. W. & Crowell, S. L., 2000 Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304. ISSN 0016-6731.

[9] Kimura, M. & Crow, J., 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

[10] Connallon, T. & Clark, G. A., 2011 The resolution of sexual antagonism by gene duplication. *Genetics* **187**, 919–937.

[11] Moran, P., 1958 Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54**, 60–71.