Supporting Information

Mustonen *et al.* 10.1073/pnas.0805909105

SI Text

Abf1 and Reb1 Energy Matrices and Site Ensembles. The energy matrices used in this study, as well as the Q_4 site ensembles, are available for download at http://www.thp.uni-koeln.de/vimuston/supp/. A README file at this link describes the formats used for the online data.

Simulation of Binding Site Evolution. Realizations of the evolution of an individual site are generated by using a standard Gillespie algorithm based on proposing point substitutions to an evolving site sequence at the neutral background rate and accepting the changes with a Kimura–Ohta fixation probability, which depends on the fitness difference between initial and proposed site sequences. The fitness difference is governed by the energies of the initial and final site sequences and by the fitness function $F(E)$, all of which we can determine as described in the main text. For the simulations reported in the main text, we estimate the neutral point substitution rate $\mu_{\mathbf{a}\to\mathbf{b}}$ from the overall statistics of point substitutions in the aligned intergenic sequence of the *cer–bay* genomes. We concentrate on point mutations because insertions and deletions (indels) play a minor role for sites of conserved function and because the neutral rates of these complex processes are hard to estimate. They are, however, important for loss-of-function events, as discussed in the main text.

We start our simulations from an ensemble of real functional sites in one genome (for Abf1, we use the 361 *cer* sites in the Q_4^{no} ensemble). For each site, we run the stochastic simulation over the yeast evolutionary tree displayed in Fig. 1 (tree topology and branch lengths as given in ref. 1) and generate a collection of quadruplets of simulated orthologous site sequences, which we wish to compare in some way with the collection of actual orthologous sites in the Q_4^{no} ensemble. The only meaningful comparisons will be statistical in nature: the connection between any individual real locus and its simulated version is loose at best. Because we have access to site energy, and are exploring the proposition that energy is the phenotype that most directly governs function, it makes most sense for us to discuss statistical measures of the variation of the energy phenotype, rather than of sequence. This has the added advantage that the available data, namely 361 sequence loci, define the statistics of a single numerical phenotype, energy, much more cleanly than the statistics of possible 14-bp binding sequences. For each locus, real or simulated, there is an energy difference ΔE (between *cer* and *par, mik* or *bay*) and, although the individual ΔE values cannot meaningfully be compared, their distribution $\Omega_f^{\tau}(\Delta E)$ (where τ indicates the appropriate divergence time) can. This distribution evolves with divergence time and it is of interest to simulate that time evolution and compare it at appropriate divergence times with the histogram of observed energy differences for the various species pairs. We generate a large number of simulated histories, starting from the same initial sites, to compute both the mean value of $\Omega_f^{\tau}(\Delta E)$ and its variance (the latter being used to assess whether the observed ΔE histogram is consistent, within finite sample noise, with the simulations). The simulations can also be used to predict the evolution of the summary quantities $\langle(\Delta E)^2\rangle$ and $\langle\delta^2\rangle$ introduced in the main text.

Inference of Functional Sites by Energy Conservation. In a first step, we align intergenic sequence of all four yeast species and require that a site in *cer* and its orthologous sites in at least two of the other species have Abf1-binding energies below an optimized cutoff $\widetilde{E} = 0.9$. We use Bayesian analysis, as in ref. 2, to estimate the remaining fraction of false positives due to neutrally evolving background loci as a function of the cutoff energy. For $\widetilde{E} = 0.9$, this fraction is $\approx$11% for the three-species sets and $<$3% for Q_4, compared with $>$50% for the single-species set Q_c. The energy distributions of the three-species sets and of Q_4 provide estimates of $Q_{\mathrm{f}}(E)$ that are quite similar to each other and to the single-species low-energy excess counts in *cer*. Because intergenic yeast DNA is subject to many functional constraints, it is also important to estimate how many loci in these sets are conserved for reasons other than selection for Abf1 binding. However, any functionality unrelated to Abf1, such as binding of other transcription factors, will constrain the sequence of a locus, but not its Abf1 binding energy E. Hence, the Abf1 energy phenotype distribution of loci unrelated to Abf1 should be $P_0(E)$, not $Q_{\mathrm{f}}(E)$. Because no admixture of $P_0(E)$ is visible in the actual energy distributions, the corresponding fraction of false positives can be at most a few percent as well.

Cutoff-Independent Statistics of Conserved Sites. To construct the ensembles Q_4, Q_4^{no}, and Q_4^{ol} of putative functional sites, we prune out spurious low-energy binding sites by imposing the requirement that the orthologs of a given site in four yeast species all have energy E less than a suitably chosen cutoff $\widetilde{E}$, as explained in the main text. Clearly, the number of sites contained in these sets, as well as the fractions of spurious sites, depend on the choice of this cutoff: as it is increased, the fraction of nonfunctional sites contained (false positives) increases and the fraction of functional sites not contained (false negatives) decreases. Biological properties of functional sites should, of course, be independent of this cutoff. We show here that the energy divergence measures $\langle(\Delta E)^2\rangle$ and $\langle\delta^2\rangle$ used in the main text are independent of the cutoff $\widetilde{E}$, evidence that the reduced value of $\langle(\Delta\mathrm{E})^2\rangle$ is not an artifact introduced by our choice of $\widetilde{E}$. Our argument is based on a careful error analysis decomposing these observables into their contributions from functional sites (which turn out to be cutoff-independent in the relevant regime) and from background sites (which are strongly cutoff-dependent). First, the normalized joint distribution of the N_4 four-species energy quadruples (E_{c}, E_{p}, E_{m}, E_{b}) can be written in the form

$$W(E_c, E_p, E_m, E_b) = \lambda_f \mathrm{Q}_f(E_c, E_p, E_m, E_b) + (1 - \lambda_f)P_0(E_c, E_p, E_m, E_b), \quad [1]$$

where λ_{f} is the occurrence fraction of functional sites, $\mathrm{Q_f}$ is the normalized component due to functional sites under selection, and P_0 is the normalized component due to neutrally evolving sequence (2). The cumulative distribution of all sites

$$\widehat{W}(\widetilde{E}\,) \equiv \int_0^{\widetilde{E}} dE_c dE_p dE_m dE_b W(E_c, E_p, E_m, E_b) \quad [2]$$

has a corresponding decomposition into cumulative distributions $\widehat{Q}_{\mathrm{f}}(\widetilde{E})$ and $\widehat{P}_0(\widetilde{E})$ of functional and background sites:

$$\widehat{W}(\widetilde{E}) = \lambda_f \widehat{\mathrm{Q}}_f(\widetilde{E}) + (1 - \lambda_f)\widehat{P}_0(\widetilde{E}). \quad [3]$$

The full cumulative distribution $\widehat{W}(\widetilde{E})$ can be estimated from the actual site data; at the same time, the cumulative distribution

$\hat{P}_0(\widetilde{E})$ of neutral sites can be estimated by stochastic simulation of the evolution process along the four-species phylogenetic tree (we use the simulation method described in the previous section, omitting the effect of energy-dependent fitness). The occurrence fraction of functional sites, λ_f, is not known *a priori* but different values will, by Eq. **3**, imply different forms for the cumulative distribution of functional sites, $\hat{Q}_f(\widetilde{E})$. Hence, we can jointly infer λ_f and $\hat{Q}_f(\widetilde{E})$ by requiring that no functional sites lie above $E > 1.2$ [equivalently, that the cumulative distribution $\hat{Q}_f(\widetilde{E})$ reaches a maximum value below this energy]. Fig. S3*a* shows the result of applying this procedure to the $\mathcal{Q}_4^{no}$ sites. The cumulative distribution $\hat{Q}_f(\widetilde{E})$ is seen to reach its asymptotic plateau already for $\widetilde{E} > 0.8$, motivating our choice of $\widetilde{E} = 0.9$ as the cutoff energy for conserved functional sites. In the entire plateau region, the inferred number of functional sites, $\lambda_f \hat{Q}_f(\widetilde{E})N_4$, and their energy distribution $Q_f(E) = d\hat{Q}_f(E)/dE$, are only weakly dependent on the actual cutoff chosen. We can also estimate the fraction of false positives, $(1 - \lambda_f)\,\hat{P}_0(\widetilde{E})/\hat{W}(\widetilde{E})$, as a function of the cutoff energy. As stated in the main text, for our cutoff choice, $\widetilde{E} = 0.9$, this fraction evaluates to $\approx 3\%$.

A similar analysis of the cutoff-dependence can be performed for the cross-species phenotype divergence $<(\Delta E)^2>$ and the linear divergence $<\delta^2>$ of conserved sites. We define the cutoff-dependent moment of the energy difference $\Delta E = E_1 - E_2$ between any two species by

$$\langle(\Delta E)^2\rangle(\widetilde{E}) \equiv \hat{W}(\widetilde{E})^{-1} \int_0^{\widetilde{E}} dE_c dE_p dE_m dE_b \times (E_1 - E_2)^2 W(E_c, E_p, E_m, E_b). \quad [4]$$

By decomposing W into its components $Q_f(E_c, E_p, E_m, E_b)$ and $P_0(E_c, E_p, E_m, E_b)$, the total variance can be written as a sum of contributions from functional and background sites:

$$\hat{W}(\widetilde{E})\langle(\Delta E)^2\rangle(\widetilde{E}) = \lambda_f \hat{Q}_f(\widetilde{E})\langle(\Delta E)^2\rangle_f(\widetilde{E}) + (1 - \lambda_f)\,\hat{P}_0(\widetilde{E})\langle(\Delta E)^2\rangle_0(\widetilde{E}). \quad [5]$$

The neutral part $\langle(\Delta E)^2\rangle_0(\widetilde{E})$ is obtained by the same stochastic simulation as used to estimate the P_0 distribution. Fig. 3 *b–e* shows, for different species pairs, the cutoff-dependent full divergence $\langle(\Delta E)^2\rangle(\widetilde{E})$ along with the contribution $\langle(\Delta E)^2\rangle_f(\widetilde{E})$ it receives from the functional sites (inferred by using Eqs. **3-5**) by using the set $\mathcal{Q}_4^{no}$ as the functional site list. The important point is that these divergence measures are independent of the choice of cutoff in the energy range $0.7 < \widetilde{E} < 0.95$ (within finite-sample errors). A similar procedure may be used to analyze the cutoff dependence of the linear divergence $\langle\delta^2\rangle$ between species (and the corresponding site-to-site variance measures within a single species). The conclusion is that these measures are also cutoff–insensitive in the energy range $0.7 < \widetilde{E} < 0.95$.

We conclude from this analysis that the evolutionary conservation of energy manifested in the behavior of these observables in the range $E < \widetilde{E}$ is an effect of the fitness landscape $F(E)$, and not an artifact of the procedure used to select our ensembles of functional sites.

Direct Comparison with Binding Assay Data. It is instructive to test the consistency of our inference procedure by direct comparison with binding assay data. As displayed in Fig. S4*a*, the histogram of ChIP-Chip binding scores (8) for 6168 promoter regions in *cer* (dots) has a peak centered around zero because of the background of nonbinding promoters and a tail at high score that is used to experimentally identify Abf1-binding promoters. The histogram of binding scores of the subset of promoters containing $\mathcal{Q}_4^{no}$ sites (bars) is strongly biased toward the high-score (binding) tail of the full distribution. The analysis of PBM binding assay data (7), shown in Fig. S4*b*, yields a similar result. These observations are a direct indication of consistency between the binding assay experiments and our energy conservation method of identifying binding sites.

Conservation of Reb1 Binding Sites. We have repeated our binding site evolution analysis and simulation for binding sites of a related TF, Reb1. The results, summarized in Figs. S1 and S2 are similar to what we found for Abf1. The impact of segregating functional sequences, selected by four-species energy conservation, into "nonoverlap" and "overlap" subpopulations is even more dramatic than for Abf1. The match between stochastic simulation of the evolution of the energy phenotype and actual site data are remarkably accurate for the nonoverlap sites, and the mismatch with "overlap" site data are striking. As was observed for Abf1, the evolution of the single-function sites is well described by a time-independent energy-dependent fitness function, whereas the multifunction sites evolve much more slowly.

Identifying Sites Overlapping with Other Factors. Functional binding sites of a given yeast TF often overlap with the binding sites of other factors, leading to more complex constraints on site sequence (3–5). For our purposes, it is useful to have a set of binding sites that are, to the largest extent possible, subject to the selection effects of Abf1 and nothing else. To address this point, we select from the $\mathcal{Q}_4$ set of sites, which are highly enriched in functional sites by energy conservation, a subset of sites further defined by having no significant overlap with binding sites of other *cer* TFs. This is done by using the collection of yeast TF position weight matrix (PWM) profiles of ref. 6. We declare an Abf1 site to have overlap if it physically overlaps with any PWM-predicted site for any other TF. A PWM site prediction is defined by requiring the score of the orthologous site sequences in all four species to be below the binding score cutoffs proposed in ref. 5. This procedure yields a set of 361 sites with no overlap, which we call $\mathcal{Q}_4^{no}$, and a complementary set of 347 sites with overlap, which we call $\mathcal{Q}_4^{ol}$. Fig. S6*a* shows that the energy distributions of the two subsets are similar, indicating that the sites with overlap are still functional Abf1 sites. A significant difference is that, as shown in Fig. 3, the putatively multifunctional sites evolve substantially more slowly than do binding sites subject to selection by a single Abf1 factor. The functional difference between these two sets of sites is also evident in the way sequence, as opposed to energy, evolves. Fig. S6 *b* and *c* show the mean probability of sequence substitutions between *cer* and *bay*, as a function of position across the Abf1 binding site and in its immediate neighborhood, for the two binding site ensembles. The mutation profiles of the two classes of site are very similar, but the mutation rate of the multifunction sites is reduced by $\approx 30\%$ across the board. Although this makes qualitative sense, it is beyond the scope of this article to quantify the selection effects that lead to this behavior.

Loss and Gain of Function. One frequently encounters cases where a below-threshold ($E < \widetilde{E}$) site in one genome is aligned to sequence in a second genome with insertions or deletions (indels). In assessing whether there is an orthologous conserved site in the second genome, we scan the energy matrix over consecutive sequence in a small window around an anchor point defined by the alignment; if there is no below-threshold site in this window, we declare this to be a loss-of-function event. To acquire data on the dynamics of loss-of-function events, we would further like to assign an energy $E' > \widetilde{E}$ to a site in the second genome into which the originally functional site was

transformed by mutations. In the absence of indels this is straightforward: we simply set E' equal to the energy of the second-genome sequence directly aligned to the below-threshold site in the first genome. In the presence of indels, this is not possible, and further heuristics are required.

A first heuristic allows us to estimate the energy change $\Delta E = E' - E$ of a loss event in the presence of indels. The presence of indels means that there are multiple ways of aligning the site in the first genome to a string of consecutive bases in the second genome and the heuristic amounts to trying the different possibilities and accepting the one with the lowest (but still above-threshold) energy. More precisely, for a given nucleotide a_i of the functional binding site **a** with aligned nucleotide a'_i in the nonfunctional site, the binding energy E'_i is defined by assuming that the TF binds with a'_i at position i of the binding sequence, and we take $E' = \min_{i=1,\ldots,\ell} E'_i$. This is a reasonable way to separate local evolutionary moves from larger rearrangements, because the number of loss events in Fig. 3 *E–G* roughly follows the corresponding external branch lengths of the phylogenetic tree. This heuristic is used to create the data displayed in Fig. 3 *e–g* on the ΔE distribution of loss-of-function events.

The fact that a loss event is associated with an indel in the alignment does not necessarily mean that the loss of function was actually because of insertion or deletion of base pairs: function could have been lost as the result of point mutations and then, after function was lost, an indel event could have occurred under neutral evolution. Because our modeling of site evolution dynamics includes only the effect of point mutations, it is important for us to try to distinguish the two cases in our census of observed loss-of-function events. This is the context for our second heuristic: a loss event is called indel-dependent if the site remains functional ($E < \widetilde{E}$) before the first indel and loses function ($E > \widetilde{E}$) because of this indel. By "before an indel" we mean that where the alignment has placed a gap in one genome, we fill in the gap with sequence from the other genome. This creates the site sequence that would have existed before a notional event that excised the bases in the gap (but after the occurrence of all of the point mutations implicit in the alignment). We can assign an energy to this before-the-indel sequence and, if that energy is below $\widetilde{E}$, we declare this to be an indel-dependent loss event. Because the initial and final site sequences **a** and **a**′ do not determine the temporal order of point substitutions and indels between them, we assume a scenario where all of the point substitutions happened first followed by the indel.

Calibration of the Neutral Mutation Model. Neutral mutation rates for intergenic sequences are difficult to estimate because they can vary across the genome and it is not known *a priori* which loci (if any) are evolving without any selective constraints. In this study, we infer these rates using a HKY model (9) calibrated in two alternative ways:

1. By using substitution counts averaged over all intergenic sequence.
2. By using the flanking bases of Abf1 sites in the Q_4^{no} ensemble as sequence for calibration. This leads to higher inferred neutral mutation rates as shown in Fig. S6*b*.

At the moment, we do not know which method is the more accurate. However, because a third method of calibrating the background model, that of using synonymous substitutions in coding regions, gives neutral substitution rates in between the two models, we think the observed differences provide a realistic estimate of the uncertainties in the inference of neutral rates. Our simulations of sequence and energy phenotype evolution reported in the main text use model 1 (see *Methods*). As expected, using model 2 leads to slightly faster phenotype evolution of binding sites compared with model 1 as displayed Fig. S7. Both models give a slight but significant underestimation of the actual evolutionary constraint on the energy phenotype: it is quite possible that the set Q_4^{no} still contains a small fraction of sites that are multifunctional for other reasons than what we have screened for.

1. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.
2. Mustonen V, Lässig M (2005) Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proc Natl Acad Sci USA* 102:15936–15941.
3. Doniger SW, Huh J, Fay JC (2005) Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res* 15:701–709.
4. Erb I, van Nimwegen E (2006) Statistical features of yeasts transcriptional regulatory code. *IEEE Proceedings of the First International Conference on Computational Systems Biology (ICCSB)*, pp 111–118.
5. Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3(5):0932–0942.
6. Harbison CT, *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.
7. Mukherjee S, *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 36:1331–1339.
8. Lee TI, *et al.* (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* 298:799–804.
9. Hasegawa M, Kishino H, Yano T-A (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.

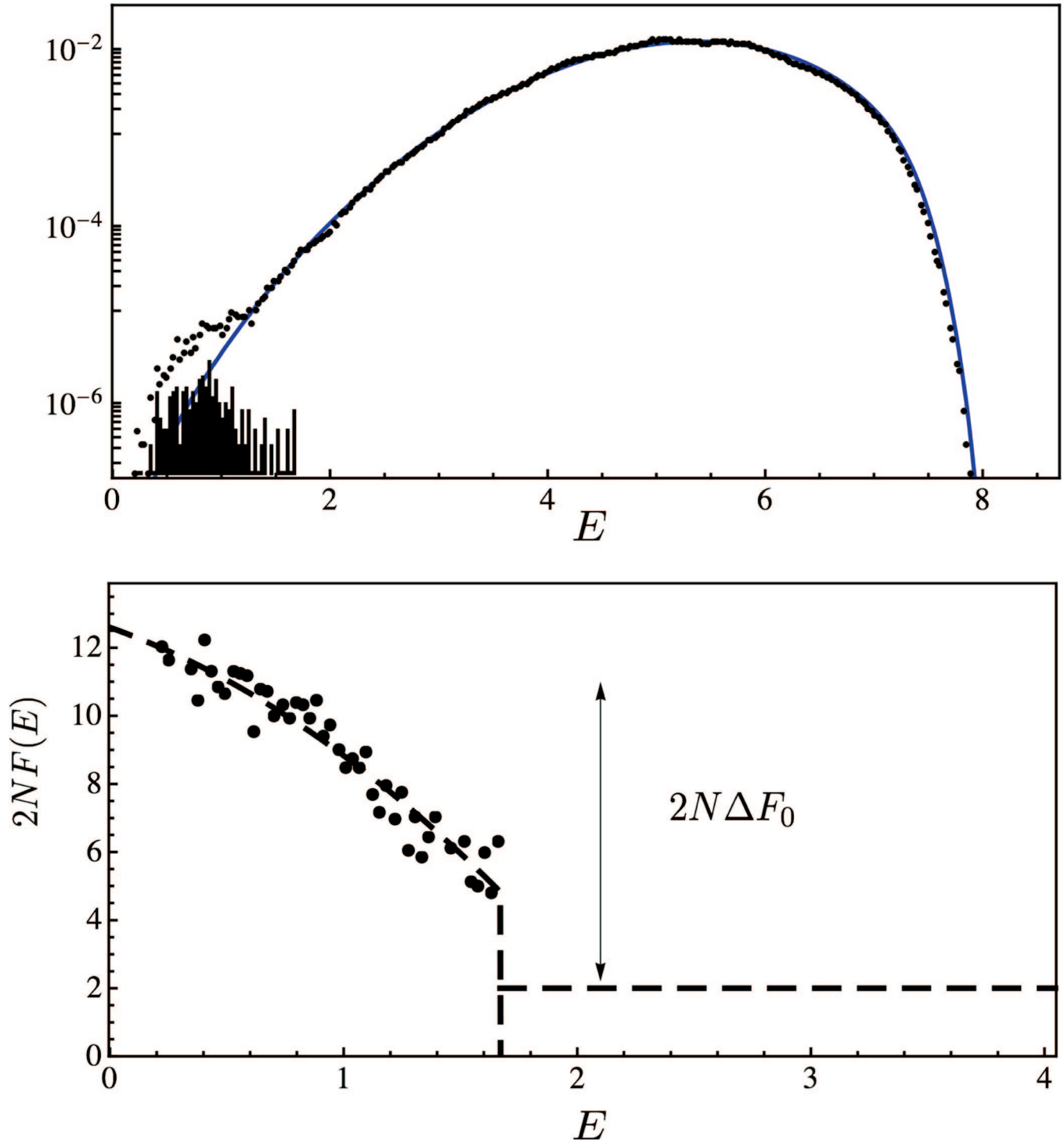

Fig. S1. Energy distribution and fitness landscape of yeast Reb1 binding sites. (*a*) Histogram of Reb1 binding energies of all sites in the intergenic part of the *S. cerevisiae* genome (points); energy histogram of the 227 Reb1 sites in the $\mathcal{Q}_4^{\text{no}}$set of nonoverlap functional sites (bars); and simulated energy distribution $P_0(E)$ of neutral background sites (solid line). The functional sites have energies below the threshold $\tilde{E} \simeq 1.6$. (*b*) Fitness landscape $F(E)$ derived from the energy distribution of $\mathcal{Q}_4^{\text{no}}$ sites and neutral distribution $P_0(E)$ according to Eq. **2** of main text (points); quadratic fit to $F(E)$ in the binding range $E < \tilde{E}$, followed by constant value in the nonbinding range $E > \tilde{E}$ (dashes). Nonbinding range ($E > \tilde{E}$): $2N\,F(E)$ is approximated as constant with difference $2N\,\Delta F_0$ to maximal binding.

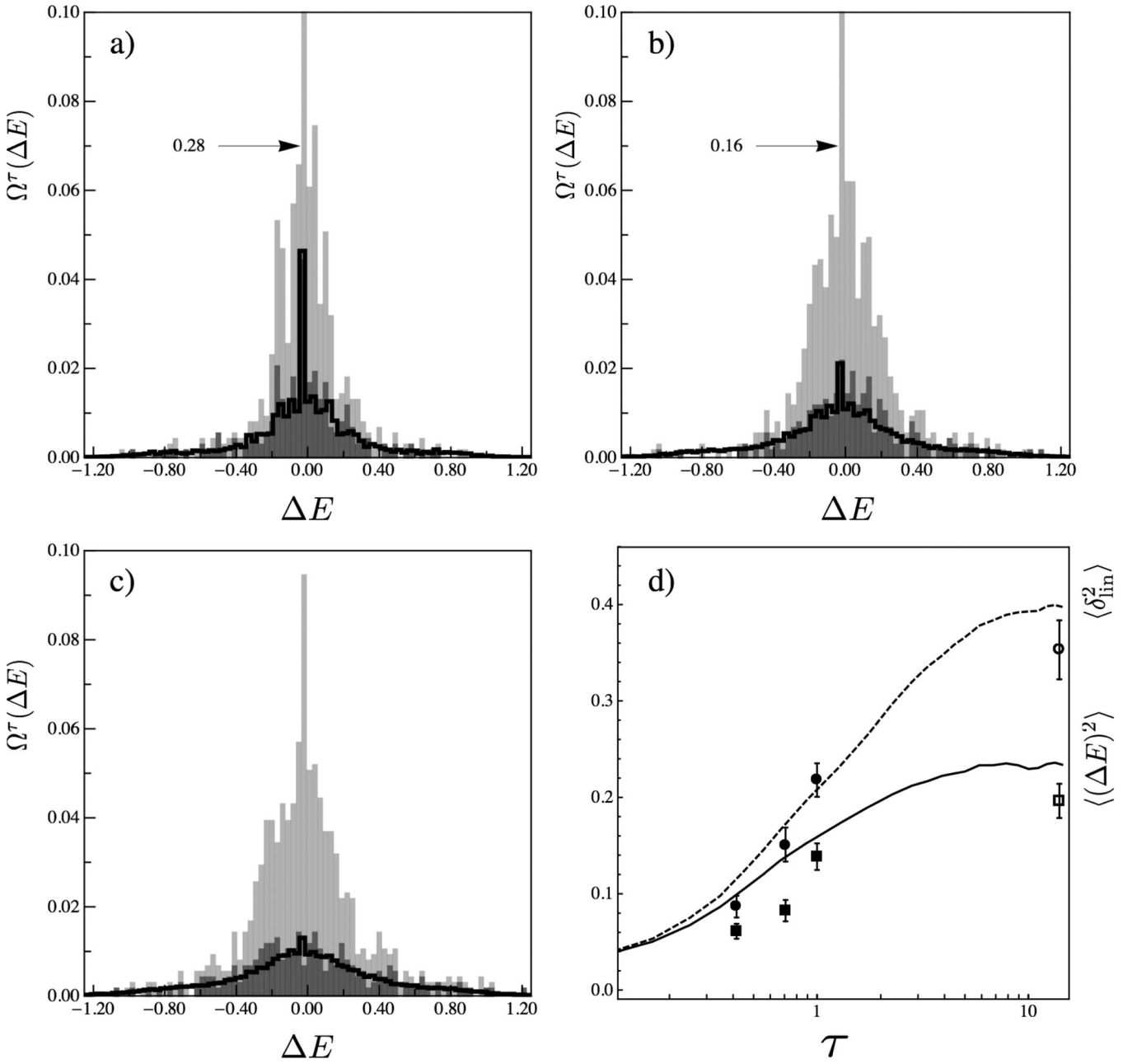

Fig. S2. Phenotype evolution of Reb1 binding sites. (*a*) Histogram of *cer–par* energy differences $\Delta E = E_{par} - E_{cer}$ for conserved Reb1 binding sites in Q_4 (bars) decomposed into contributions of nonoverlapping sites (dark gray) and overlapping sites (light gray). Predicted distribution $\Omega^\tau(\Delta E)$ (line, normalization matching the counts of nonoverlapping sites). (*b* and *c*) Same as *a* for *cer–mik* and *cer–bay* energy differences. (*d*) Energy divergence $\langle(\Delta E)^2\rangle$ (solid squares) and additive divergence δ^2 (dots; see text) for conserved nonoverlapping sites between *cer* and any of the other three species, plotted against evolutionary distance τ; predicted energy divergence $\langle(\Delta E)^2\rangle(\tau)$ (solid line) and additive divergence $\delta^2(\tau)$ (dashed line). The large-τ limit of these functions reproduces (up to sampling effects) the site-to-site energy variance in *cer* (open square) by definition of the evolution model and predicts the site-to-site linear variance (open dot).

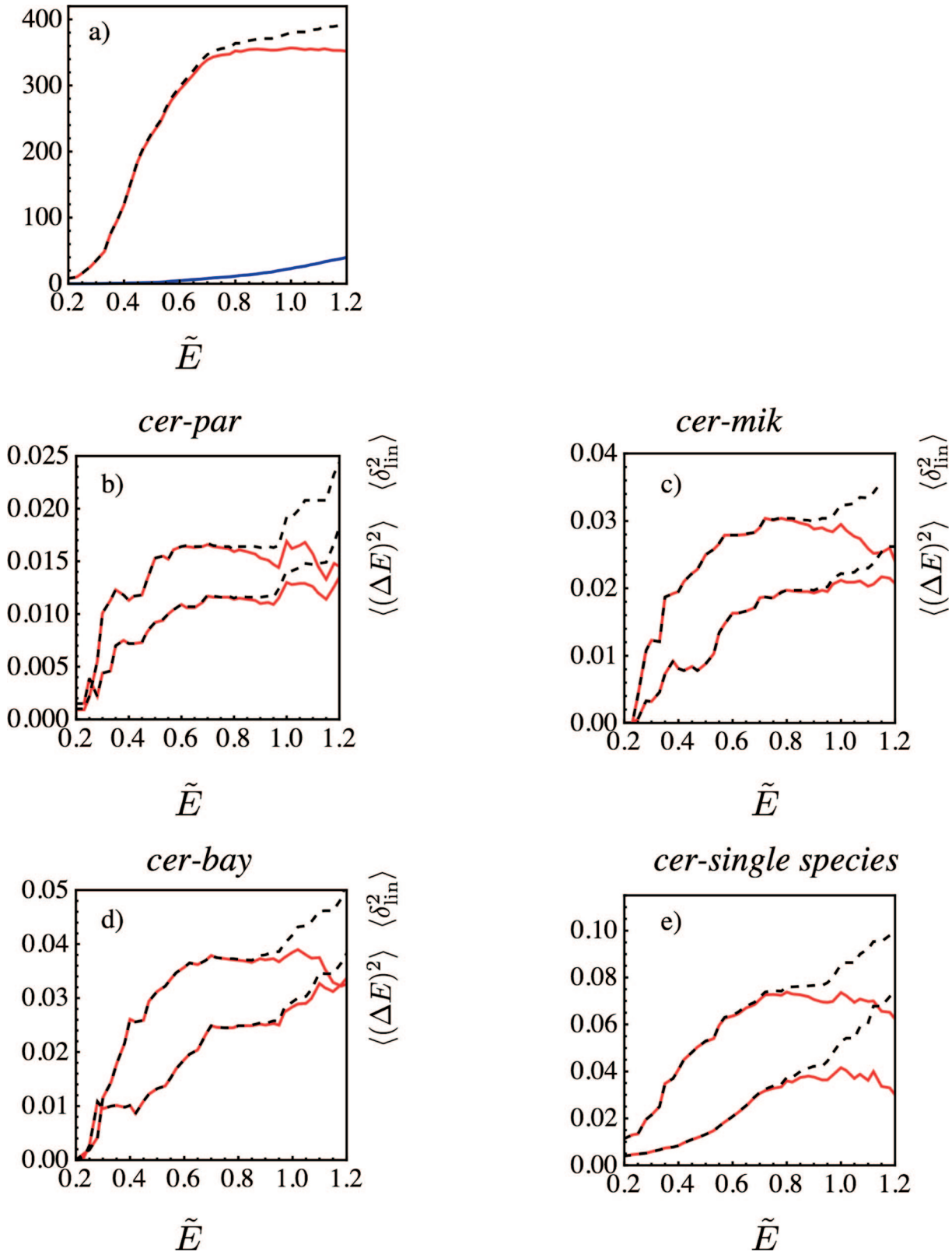

Fig. S3. Cutoff-dependent statistics of conserved sites. (*a*) Cumulative distribution $N_4\,\hat{W}(\tilde{E})$ of $Q_4^{\rm no}$ sites with energies $E < \tilde{E}$ in all four species (dashed line) and its decomposition into the contribution $N_4\,\hat{P}_0(\tilde{E})$ of neutrally evolving false positives (blue line) and the contribution $N_4\,\hat{Q}_f(\tilde{E})$ of functional sites under selection for Abf1 (red line). (*b*–e) Cutoff-dependent phenotype divergence $\langle(\Delta E)^2\rangle$ ($\tilde{E}$), linear divergence $\langle\delta^2\rangle$ ($\tilde{E}$) (dashed lines), and the inferred contributions $\langle(\Delta E)^2\rangle_f$ ($\tilde{E}$) and $\langle\delta^2\rangle_f$ ($\tilde{E}$) for different species pairs (red line).

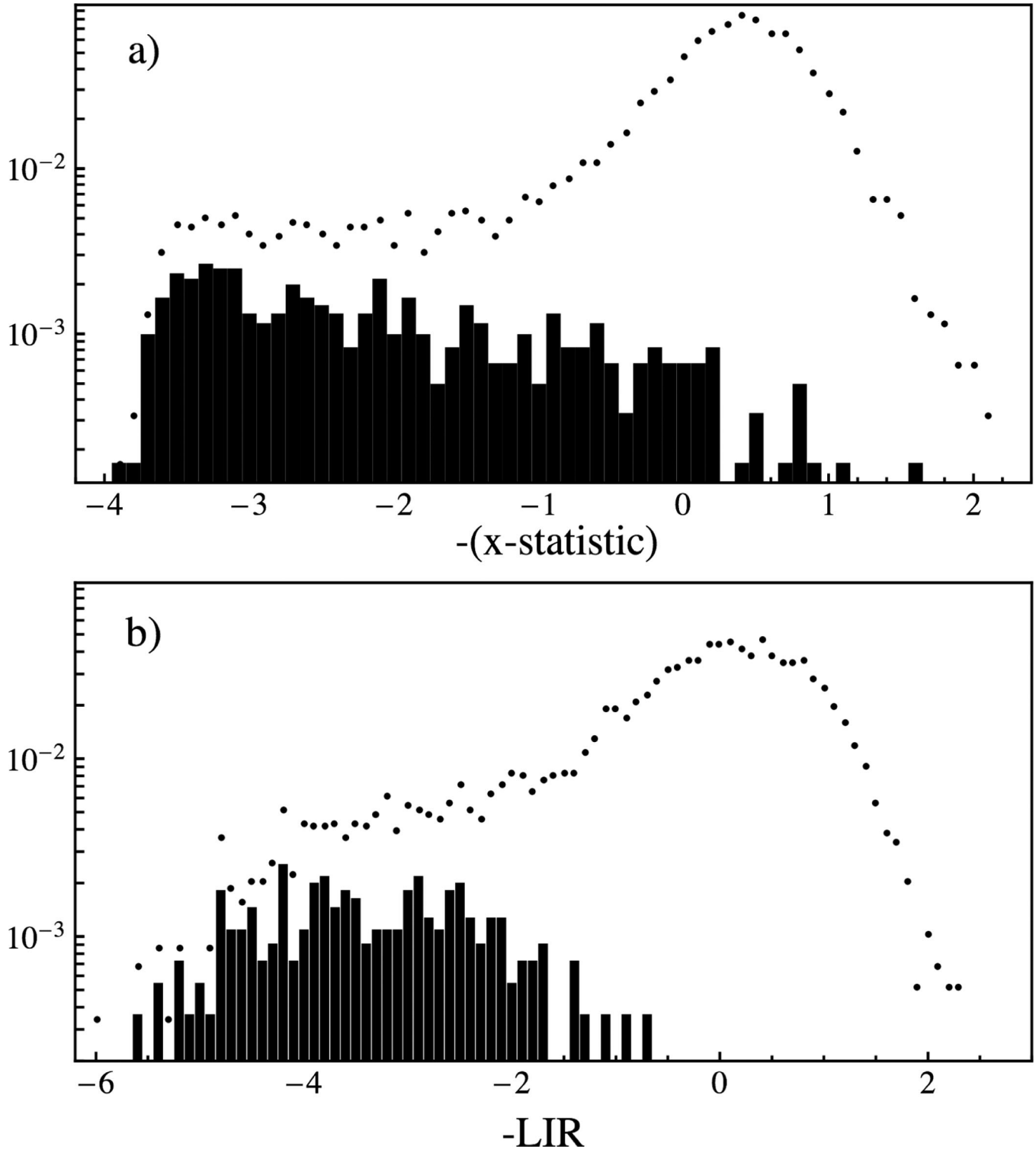

Fig. S4. Histograms of Abf1 binding assay data for all *cer* promoters (dots) and for the subset of promoters which contain at least one Q_4^{no} site (bars). (*a*) ChIP-chip data (7) (LIR, observed log-intensity ratio). (*b*) PBM data (8) (*x* statistic, statistically filtered measure of fluorescence log-intensity ratios). The promoters with Q_4^{no} sites are strongly biased toward the binding tail of the distribution of all promoters, indicating consistency between our analysis based on evolutionary conservation of energy and the experimental binding assays.

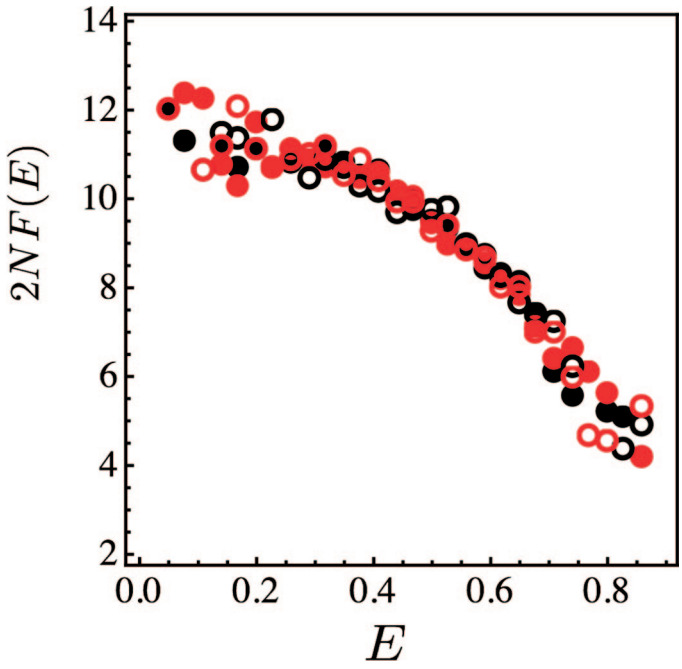

Fig. S5. Species-specific fitness landscapes for Abf1 inferred from binding site energy distributions in four species: *cer* (black-filled circles), *par* (red-filled circles), *mik* (black open circles), and *bay* (red open circles). The landscapes $F(E)$ are very similar, consistent with the assumption that the Abf1 loci in all four species are close to phenotypic equilibrium in a common fitness landscape.

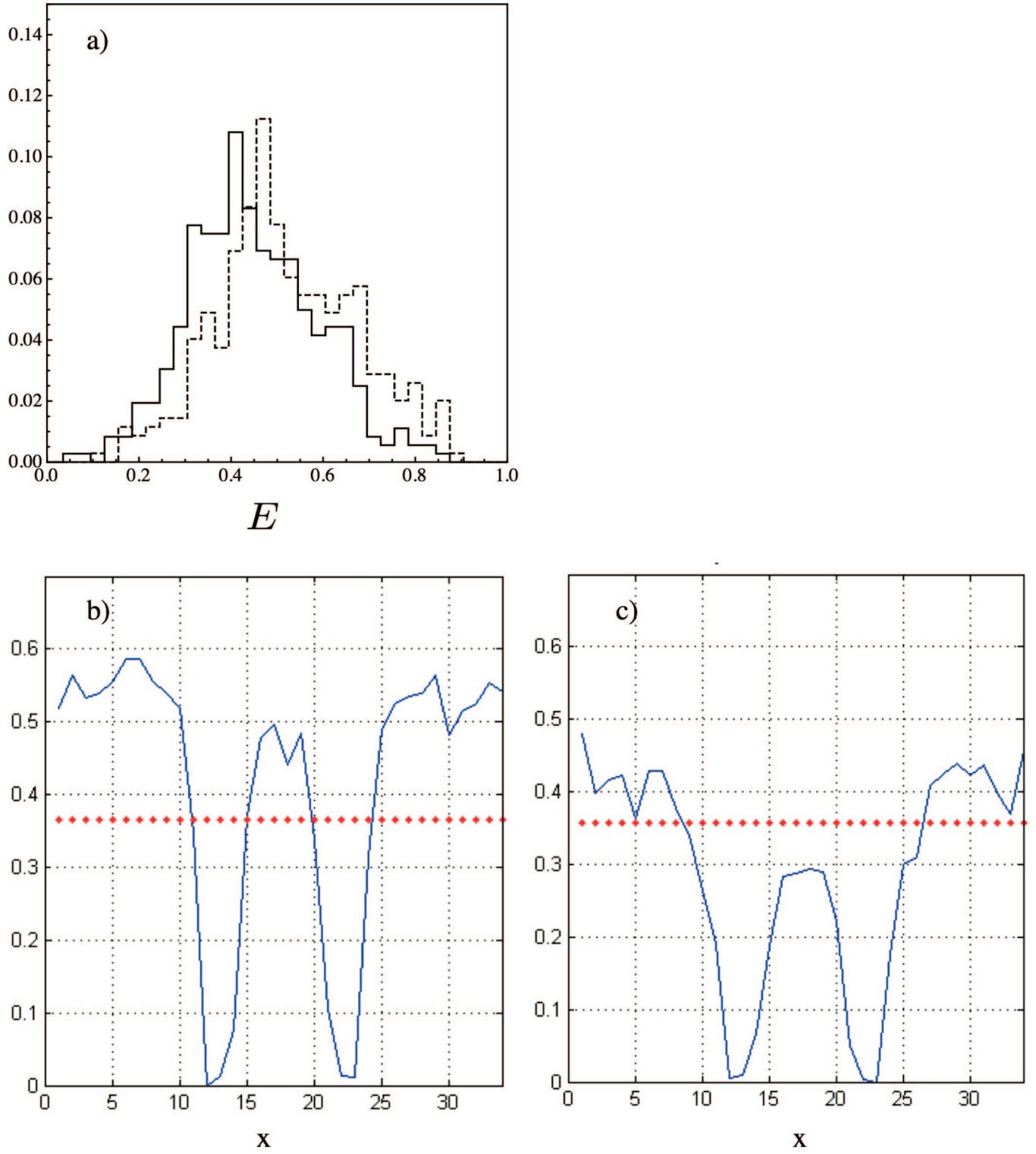

Fig. S6. Multifunctional sequences. (*a*) $Q_f(E)$ distributions for sites without overlap with other TF binding sites (solid line) and with overlap (dashed lines). The distributions are similar and consistent with the fitness landscape (see Fig. 3) acting on the Abf1 binding site phenotype. (*b* and *c*) The substitution rate profiles of Abf1 *cer–bay* site pairs. Ten base pairs of flanking intergenic positions on either side of the core motif (which starts at $x = 10$) are included; dots indicate mean single-point substitution rate across all aligned intergenic sequence. (*b*) Mutation profile derived from the $\mathcal{Q}_4^{\mathrm{no}}$ set (selected for no overlap with other TF sites). (*c*) Mutation profile derived from the overlap set $\mathcal{Q}_4^{\mathrm{ol}}$; additional constraints on the sequence reduce the mutation rate within the site and in the flanking positions.

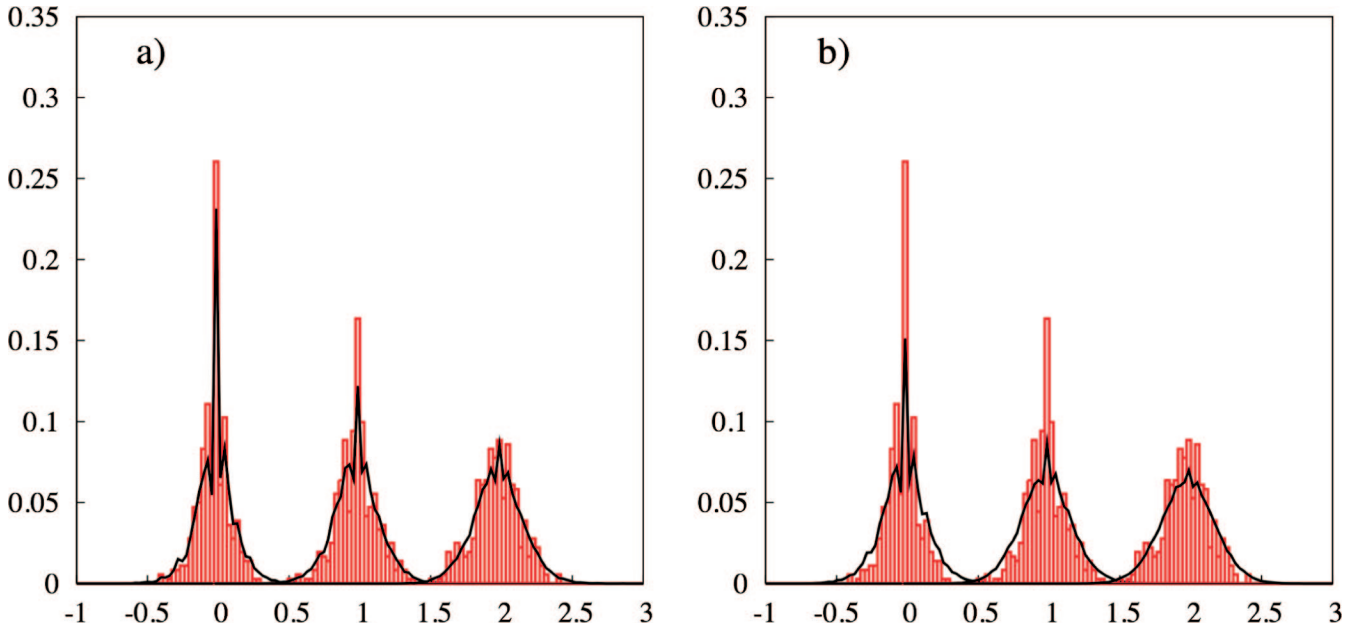

Fig. S7. Testing the effect of the background mutation model calibration. ΔE histogram for actual orthologous site pairs (red bars); stochastic simulation of the same data (black curve). Shown, *Left* to *Right*, results for the sequence pairs *cer–par*, *cer–mik*, *cer–bay* (with ΔE abscissa displaced for clarity by 0.0, 1.0, and 2.0, respectively). (*a*) Stochastic simulation with inferred $F(E)$ and HKY model based on all intergenic sequence. (*b*) Simulation with $F(E)$ and HKY model based on intergenic sequence flanking the nonoverlap sites Q_4^{no}. Real data (red bars) are the same in both *a* and *b*. As discussed in *SI Text*, the latter neutral model evolves somewhat more rapidly than the former. Outside the $\Delta E \sim 0$ peak, the two models are nearly indistinguishable.