Bioenergetic costs and the evolution of noise regulation by microRNAs

Efe Ilker[1] and Michael Hinczewski[2]

[1] *Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*
[2] *Department of Physics, Case Western Reserve University, Cleveland OH 44106*

Noise control, together with other regulatory functions facilitated by microRNAs (miRNAs), is believed to have played important roles in the evolution of multicellular eukaryotic organisms. miRNAs can dampen protein fluctuations via enhanced degradation of mRNAs, but this requires compensation by increased mRNA transcription to maintain the same expression levels. The overall mechanism is metabolically expensive, leading to questions about how it might have evolved in the first place. We develop a stochastic model of miRNA noise regulation, coupled with a detailed analysis of the associated metabolic costs. Additionally we calculate binding free energies for a range of miRNA seeds, the short sequences which govern target recognition. We argue that natural selection may have fine-tuned the Michaelis-Menten constant K_{M} describing miRNA-mRNA affinity, and show supporting evidence from analysis of experimental data. K_{M} is constrained by seed length, and optimal noise control (minimum protein variance at a given energy cost) is achievable for seeds of 6-7 nucleotides in length, the most commonly observed types. Moreover, at optimality the degree of noise reduction approaches the theoretical bound set by the Wiener-Kolmogorov linear filter. The results illustrate how selective pressure toward energy efficiency has potentially shaped a crucial regulatory pathway in eukaryotes.

Nonequilibrium processes within living systems exact a high price: the constant maintenance of fuel molecules and raw materials at sufficient concentrations to provide thermodynamic driving potentials for biological function [1]. Optimizing that function with respect to thermodynamic costs is a factor constraining evolution, and would have been particularly important at the very earliest stages of life where the metabolic chemistry responsible for maintaining those potentials was necessarily primitive and relatively inefficient. Yet thermodynamic costs are not the only factor that matters, and biology is full of counter-intuitively complex chemical mechanisms whose evolutionary predecessors, perhaps arising out of the randomness of genetic drift, may have consumed energy resources without any clear fitness benefit.

The discovery of microRNAs (miRNAs) along with their counterparts, small non-coding RNAs, raised many open questions about their functional purposes and evolution [2, 3]. These short endogenous RNAs, around 22 nucleotides (nt) in length, exist in many eukaryotic cells. They constitute the core of the RNA-induced silencing complex (RISC) that interacts with target messenger RNA (mRNA), leading to translational repression and the accelerated degradation of their target by a mechanism known as the RNA interference. One possible functional role for this interference, which is the focus of our work, is fine-tuning noise in protein populations by reducing the variance of protein copy numbers [4, 5], conferring robustness to cellular functions [6]. Such noise control, together with other regulatory functions facilitated by microRNAs, is believed to have played important roles in the evolution of complex multi-cellular life [7–10]. Yet it is a considerable expenditure of resources, similar to setting up a factory production line for a valuable good, funding gangs of thieves to constantly raid the factory, and compensating for losses by increasing the production rate. *So how would such a regulation scheme arise, and has evolution actually optimized it?* Using a combination of statistical physics, information theory, biochemistry, and population genetics, we arrive at some tentative answers to these questions.

Our results give insights into one peculiar feature of this system: why the job requires such short RNA molecules—the significance of the "micro" in microRNA. Specific interactions between the miRNA and its target in fact largely depend on only a 6-8 nt sequence known as the miRNA seed region, which forms Watson-Crick base pairs with a complementary sequence on the target mRNA. We argue that the observed length of the seed region lies in a metabolic sweet spot, giving just enough affinity between the miRNA and mRNA (measured by their Michaelis-Menten constant K_{M}) to optimally control noise at a given level of energetic expenditure. Longer seed sequences (higher affinity, smaller K_{M}) would increase rather than decrease noise. On the other hand, shorter seed sequences (lower affinity, larger K_{M}), while allowing interactions with a wider range of targets, would also require significantly more energy to achieve the same level of control. By estimating these energy expenditures, we also show that effective noise control is costly enough to be under selection pressure in eukaryotic cells. A novel miRNA may initially appear during the course of evolution as a random product of genetic drift with non-optimal parameters, and then get gradually repurposed as a noise control mechanism that confers a fitness advantage. Once this starts to occur, our theory predicts that natural selection would hone K_{M} toward an optimal value.

miRNA regulation as a noise filter for protein expression. We start our theoretical description by introducing the architecture of the miRNA-regulated system $\mathcal{S}_R$ (Fig. 1B) in comparison to the unregulated one $\mathcal{S}_0$ (Fig. 1A). Our description builds off the experimentally-

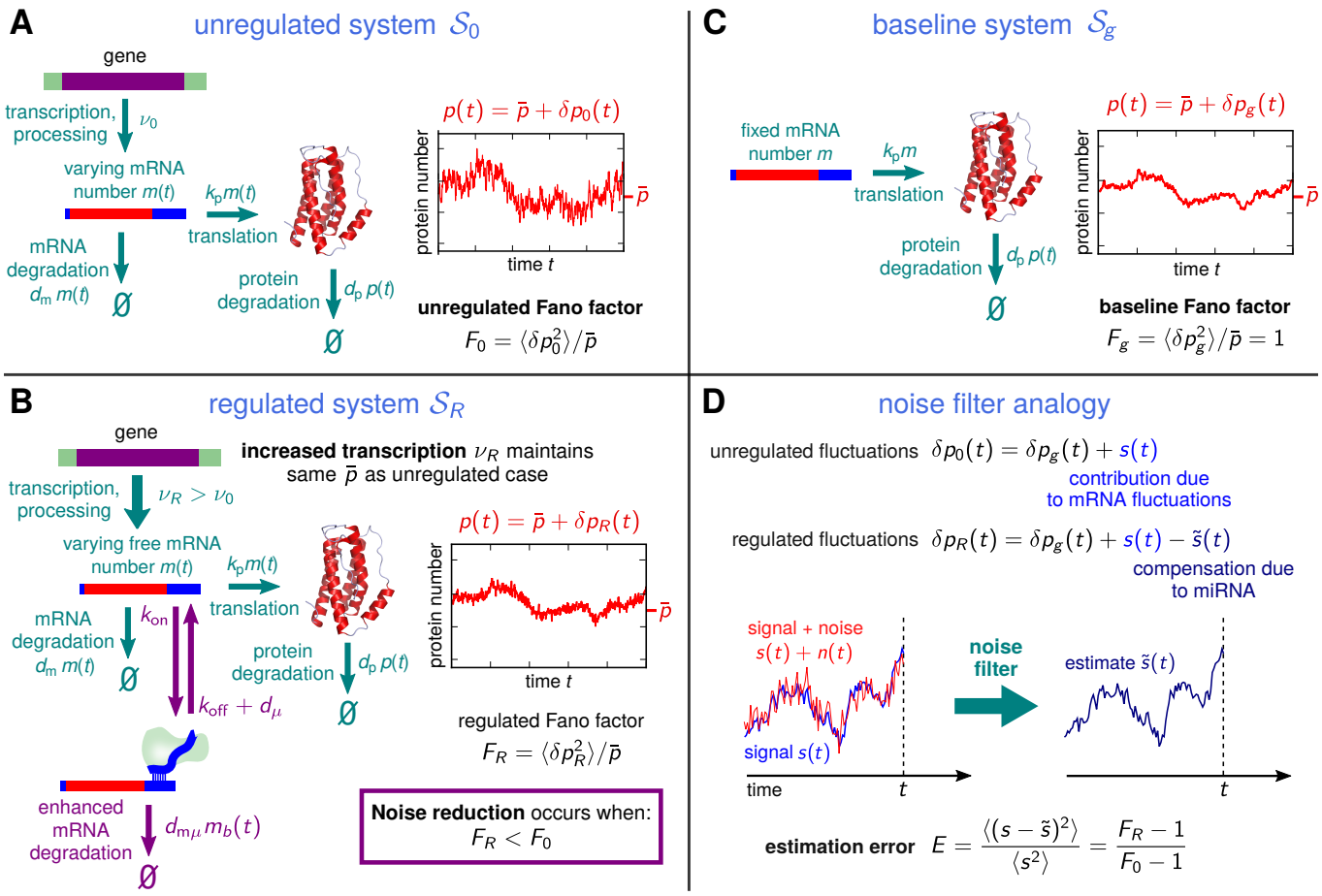

FIG. 1. **Overview of the miRNA noise regulation model:** (**A**) In the unregulated system $\mathcal{S}_0$ with no miRNA, noise in protein numbers has contributions from both varying mRNA population and intrinsic noise in the translation process, leading to a Fano factor F_0 (protein variance divided by the mean protein number $\bar{p}$). (**B**) For the regulated system $\mathcal{S}_R$, enhanced mRNA degradation due to targeting by miRNA can reduce protein noise, leading to a Fano factor $F_R < F_0$. To compensate for the loss of mRNA and achieve the same $\bar{p}$, the transcription rate ν_R must be increased relative to its value ν_0 in $\mathcal{S}_0$. (**C**) To set up the noise filter analogy, we introduce an imaginary basline system $\mathcal{S}_g$, which has fixed mRNA population and hence the protein noise is solely due to translation. The resulting Fano factor $F_g = 1$ is a lower bound on the two systems above: $1 \leq F_0$, F_R. (**D**) Decomposing the unregulated/regulated protein fluctuations $\delta p_0(t)$ and $\delta p_R(t)$ into a baseline contribution $\delta p_g(t)$ and an additional contribution allows us to define the signal $s(t)$ and estimate $\tilde{s}(t)$ in the noise filter analogy. Their normalized mean-squared difference E is the estimation error of the filter, which can be expressed in terms of F_R and F_0.

validated model of Ref. [5], and takes the form of stochastic biochemical reaction networks with dynamics obeying linearized chemical Langevin (CL) equations [11, 12], as detailed in Supplementary Information (SI) Sec. S.I A. The CL results agree closely with available experimental data [5], and give us analytical expressions for correlation functions describing copy number fluctuations of chemical species in each system.

The fluctuating output numbers are denoted as $x(t) = \bar{x} + \delta x(t)$ where $\delta x(t)$ is the variation from the mean level $\bar{x}$. The main species of interest are total (bound + unbound) miRNA, mRNA and protein copy numbers, denoted by $x = \mu_{\rm tot}$, m or p respectively. In the unregulated system $\mathcal{S}_0$ (Fig. 1A) we have transcription from a gene at rate ν_0 producing an mRNA population $m(t)$, which is then translated with rate constant k_p to a produce a protein population $p(t)$. The mRNA and proteins are degraded with rate constants d_m and d_p respectively.

To meaningfully compare $\mathcal{S}_R$ to $\mathcal{S}_0$, we assume parameter sets such that the mean protein output $\bar{p}$ is the same for both regulated and unregulated systems. This maintains the functional effectiveness of the protein in the regulated system, but with the potential added benefit of noise reduction. Indeed miRNA are often upregulated along with their target mRNA via feedforward loops [13, 14]. For $\mathcal{S}_R$ (Fig. 1B) there is an additional RNA interference mechanism: free miRNAs with population $\mu(t)$ can bind to the mRNA with rate constant $k_{\rm on}$ to form a bound complex with population $m_b(t)$. Considering a single miRNA binding site on the target mRNA, the total number of miRNA is $\mu_{\rm tot}(t) = \mu(t) + m_b(t)$. The mRNA in the complex has an enhanced degradation rate

constants $d_{m\mu} > d_m$ relative to the regular mRNA value d_m. The miRNA unbinds with rate constant $k_{\rm off}$, and degrades with d_μ. For simplicity we assume the miRNA degradation rate constant d_μ is the same when both bound and unbound to mRNA [5]. miRNA-mRNA affinity can be characterized through the Michaelis-Menten constant,

$$K_{\rm M} = K_{\rm D} + k_{\rm cat}/k_{\rm on}, \tag{1}$$

where the dissociation constant $K_{\rm D} = k_{\rm off}/k_{\rm on}$ and $k_{\rm cat} = d_{m\mu} + d_\mu$ is an effective catalytic rate constant for the miRNA-catalyzed degradation reaction. $K_{\rm M}$ approximately relates $m_b(t)$ to $\mu_{\rm tot}(t)$ through $m_b(t) \approx m(t)\mu_{\rm tot}(t)/(K_{\rm M} + m(t))$. Note that experimental values of $K_{\rm M}$ and $K_{\rm D}$ are reported in units of concentration (molars) while our CL formalism uses copy numbers of chemical species. We assume a typical eukaryotic cell volume $V = 2000\ \mu{\rm m}^3$ to convert between concentrations and copy numbers as needed.

Because of interference from the miRNA, the transcription rate ν_R of mRNA in $\mathcal{S}_R$ must be larger than the rate ν_0 in $\mathcal{S}_0$, if both systems maintain the same mean protein level $\bar{p}$. In the limit of no miRNA ($\mu_{\rm tot} \to 0$) or vanishing affinity ($K_{\rm M} \to \infty$) the regulated system approaches the same behavior as the unregulated one, with $\nu_R \to \nu_0$. The strength of regulation can be characterized by a parameter $R \equiv 1 - \nu_0/\nu_R$, where R ranges between 0 (no regulation) to 1 (maximum possible regulation). For known miRNA-mediated regulation networks, R typically lies between 0.05-0.95 [15, 16].

To quantify the effect of miRNA on noise, we look at the Fano factor $F_\lambda = \langle(\delta p_\lambda)^2\rangle/\bar{p}$ with $\lambda = R, 0$ labeling the system $\mathcal{S}_\lambda$ in which the quantity is calculated. Successful noise reduction implies that $F_R < F_0$: for the same mean protein output level, there is less protein variance in the presence of miRNA. Qualitatively this arises because the miRNA system reduces the number of translated proteins per mRNA, on average by a factor of $1-R$, and hence decreases the susceptibility of translation to fluctuations in mRNA levels. When miRNA regulation is compensated for by transcriptional increase, it is thus possible to mitigate the propagation of noise from mRNA to protein numbers. However, there is a trade-off, because the stochasticity of miRNA-mRNA interactions, as well as fluctuating miRNA populations, also introduces noise into mRNA levels. This added noise can cancel out the protein noise reduction benefit in certain parameter regimes, for example when miRNA-mRNA affinities are high.

To understand the role of miRNA more precisely, it is helpful to use a noise filter analogy. In order to motivate this mathematical analogy, we define an imaginary baseline system $\mathcal{S}_g$ (Fig. 1C), where we have fixed the mRNA population at a constant level $m(t) = \bar{m} = (d_p/k_p)\bar{p}$, which agrees with the mean $\bar{m}$ in $\mathcal{S}_R$ and $\mathcal{S}_0$ and hence gives the same $\bar{p}$. This removes the contribution of mRNA fluctuations to the noise, so $p(t) = \bar{p} + \delta p_g(t)$, where $\delta p_g(t)$ are the ground-level (baseline) fluctuations that come from protein translation and degradation (and cannot be mitigated by RNA interference). As summarized in Fig. 1D, the protein fluctuations in both $\mathcal{S}_0$ and $\mathcal{S}_R$ can then be compared to this baseline. For $\mathcal{S}_0$ we write $\delta p_0(t) = \delta p_g(t) + s(t)$, where $s(t)$ represents the added noise due to varying mRNA population $m(t)$. For $\mathcal{S}_R$ this added noise is partially compensated, $\delta p_R(t) = \delta p_g(t) + s(t) - \tilde{s}(t)$. Both $s(t)$ and $\tilde{s}(t)$ can be explicitly calculated using the CL formalism (SI Sec. S.II) and it turns out they are correlated: $\tilde{s}(t)$ takes the form of a convolution,

$$\tilde{s}(t) = \int_{-\infty}^{t} dt' H(t-t')(s(t') + n(t')), \tag{2}$$

where $H(t)$ and $n(t)$ are functions of the biochemical parameters. The above equation has a clear noise filter interpretation (depicted schematically in Fig. 1D): $s(t)$ is the "signal", $n(t)$ a "noise" that corrupts the signal, $H(t)$ is a linear filter function that acts via convolution on the past history of the corrupted signal $s(t) + n(t)$, and $\tilde{s}(t)$ is the "estimate" of the signal. It turns out the problem of reducing protein noise via mRNA interference (making F_R as small as possible) is equivalent to making $\tilde{s}(t)$ as close as possible to $s(t)$. We can see this directly in the error of estimation, which is defined as $E = \langle(s - \tilde{s})^2\rangle/\langle s^2\rangle$. For our case, E can be expressed in terms of the Fano factors, $E = (F_R - 1)/(F_0 - 1)$. Fine-tuning miRNA parameters only affects F_R, leaving F_0 fixed, so E can be minimized by decreasing F_R. Both F_R and F_0 are bounded from below by the Fano factor $F_g = 1$ of the baseline system, so perfect filtering, $E \to 0$, would correspond to $F_R \to 1$. The noise filter interpretation, which has been earlier applied to a variety of other biological networks (for a review see Ref. [17]), has an important payoff which we will return to later: it allows us to find the conditions for optimal noise reduction, and calculate tighter bounds on F_R, which in general will be greater than 1 at optimality.

Bioenergetic costs of miRNA regulation. The second major component of our model is an estimate of the costs for miRNA regulation, which we adapt from experimental data on eukaryotic transcription energetics collected in Ref. [18]. In general this includes energy expenditures channeled to the synthesis of new molecules, as well as maintenance, the recycling/repair of molecules to maintain steady-state levels (i.e. assembling mRNA from existing nucleotides to counterbalance degradation). Eukaryotic cells typically have long enough generation (cell division) times t_r, that maintenance is the dominant contribution to metabolic expenditures over a generation. Focusing on the maintenance costs, we can estimate the total metabolic consumption C_T (in units of phosphate [P] bonds hydrolyzed, namely ATP or ATP-equivalents) for the unregulated system per generation: $C_T \simeq t_r M_\nu$, where $M_\nu = \nu_0 \epsilon_m$ is the consumption rate. Here ν_0 is the mRNA transcription rate in $\mathcal{S}_0$, and ϵ_m is the energy cost in terms of P for assembling the mRNA (which will

depend on the length of the transcript). For the regulated case, the total cost will be $C_T + \delta C_T$, where the extra contribution $\delta C_T = t_r \Delta M_\nu$. The difference in consumption rate $\Delta M_\nu = (\nu_R - \nu_0)\epsilon_m + \nu_\mu \epsilon_\mu$. The first term accounts for the added costs of increased transcription to maintain the same $\bar{p}$, while the second term is the rate of miRNA assembly ν_μ times the cost ϵ_μ of that assembly in units of P (including potentially any related costs of the RISC complex). As shown in SI Sec. S.I C, ΔM_ν can be expressed in terms of the biochemical parameters of the system as

$$\Delta M_\nu = \frac{R}{1-R}\left(1 + \gamma_\mu \left(1 + \frac{K_\mathrm{M}}{\bar{m}}\right)\sigma_\epsilon\right) M_\nu, \quad (3)$$

with $\gamma_\mu \equiv d_\mu / d_{m\mu}$ and $\sigma_\epsilon \equiv \epsilon_\mu / \epsilon_m$. Based on experimental estimates (summarized in SI Table S1) we know that $\gamma_\mu \ll 1$ and $\sigma_\epsilon \ll 1$. Given the experimental range of $R = 0.05 - 0.95$, the term $R/(1-R)$ can vary by a factor of 361 between the smallest and largest observed regulation magnitudes, highlighting the strong dependence of ΔM_ν on R. The remaining terms in Eq. [3] encapsulate the modification to the costs due the parameters governing miRNA-mRNA interactions, particularly K_M and the degradation enhancement $d_{m\mu}$.

As discussed later in the section on evolutionary pressure, it is convenient to define a non-dimensional measure for the extra cost due to regulation: the extra cost as a fraction of the total metabolic expenditure of a cell per generation, $\delta C_T / C_T$. Here $C_T \simeq t_r M_\mathrm{tot}$, where M_tot is the total maintenance ATP consumption rate. Based on the data from Ref. [18], M_tot approximately scales with cell volume, and we use a value $M_\mathrm{tot} = 3 \times 10^{11}$ P/hr characteristic of eukaryotic cells. Thus we will report costs in terms of $\delta C_T / C_T = \Delta M_\nu / M_\mathrm{tot}$, with ΔM_ν given by Eq. [3].

Our model so far has assumed one mRNA target, but in general a single miRNA can target up to hundreds of mRNAs, which will also change the energetic costs of regulation. As a rough estimate of this multi-target scenario, we can assume similar biochemical parameters among different targets. This allows us to use the single-target theory but scaling up $\bar{m}$ (and hence $\bar{p}$) to reflect the total mRNA numbers when accounting for all targets. Note that the dependence of ΔM_ν on $\bar{m}$ is non-trivial, since both R and M_ν depend on $\bar{m}$. However when we demand a certain level of overall noise control (a specific value of E) the extra cost ΔM_ν will increase with $\bar{m}$ as the number of targets gets larger: the larger the system, the more expensive it is to control.

Seeds of length $6-7$ nt are most energetically efficient for noise reduction. With all the components of the model defined, we can now investigate the interplay between noise reduction and energetic costs. The error E (or equivalently the Fano factor F_R) can be expressed as a function of ΔM_ν, γ_μ, σ_ϵ, $\bar{m}$ and K_M (details in the SI Secs. S.I B,C). For a given cost ΔM_ν and fixed degradation/energy parameters γ_μ and σ_ϵ, we can ask what value of K_M minimizes E. K_M is an interesting tuning parameter because it is related to the binding strength between the miRNA and the mRNA, which in turn depends on the number of complementary interactions between the seed and the target region of the mRNA. From Eq. [1], $K_\mathrm{M} \geq K_\mathrm{D}$, and the dissociation constant K_D is related to the free energy of binding via $\Delta G^0 = k_B T \ln(K_\mathrm{D}/[1\mathrm{M}])$. Longer seeds should allow for more negative ΔG^0 (stronger binding) and hence smaller values of K_D. This in turn gives access to smaller values of K_M. For RNA interference systems, experimentally measured K_M values have ranged from comparable to K_D to about two orders of magnitude larger than K_D [19].

In Fig. 2A, we show the contour diagram of $\log_{10} E$ as a function of K_M and fractional metabolic cost $\delta C_T / C_T$, assuming a single mRNA target and using the experimentally-derived parameters of SI Table S1. The blue curve denotes the optimal value K_M^*, which achieves the minimum E for a given cost $\delta C_T / C_T$. The red contour line marks the value K_M^d where $E = 1$, which corresponds to $F_R = F_0$. This is the boundary between the noise control region to the right, where $E < 1$ ($F_R < F_0$) and a "dud" region to the left, where $E > 1$ ($F_R > F_0$). In the latter region regulation adds protein noise to the system rather than mitigating it which can provide an alternative role for some microRNA systems in triggering cell state transitions [20, 21]. For a fixed $\delta C_T / C_T$, as we scan K_M from small to large values, we cross from dud to noise control at K_M^d, improve the filter performance until we reach K_M^*, and then get progressively worse filtering for $K_\mathrm{M} > K_\mathrm{M}^*$. The different behaviors of the system with varying K_M reflect the tradeoff due to miRNA-mRNA affinity mentioned earlier in the noise filter discussion. The optimal affinity K_M^* is a metabolic sweet spot between a regime where miRNA-mRNA interaction is too strong (small K_M), leading to excessive added noise and the "dud" scenario, and a weak interaction regime (large K_M) where the miRNA system cannot effectively dampen noise.

In the biologically relevant parameter regime γ_μ, σ_ϵ, $\phi \ll 1$, where $\phi \equiv d_p / d_m$, we can derive analytical approximations for both K_M^d amd K_M^*. For small costs $\delta C_T / C_T$ we have $K_\mathrm{M}^d \approx \bar{m}(1 + \gamma_\mu^{-1})$, and with increasing cost the boundary begins to decrease as $K_\mathrm{M}^d \approx \bar{m}((1-R)/(R\gamma_\mu))^{1/2}$. On the other hand, K_M^* is remarkably stable as $\delta C_T / C_T$ is varied. In the large cost limit it approaches

$$K_\mathrm{M}^* \approx \bar{m} \gamma_\mu^{-1} \sigma_\epsilon^{-1/2}. \quad (4)$$

Thus the optimal affinity depends on two non-dimensional biochemical ratios, γ_μ and σ_ϵ, and the mean mRNA target number $\bar{m}$.

Fig. 2B shows what happens when we scale up $\bar{m}$ by a factor of 100, roughly mimicking the case of 100 similar mRNA targets. As predicted by Eq. [4], K_M^* increases by a factor of 100, but the contour levels are pushed up by a similar factor: as expected, it costs more achieve the same level of noise control when compared to the

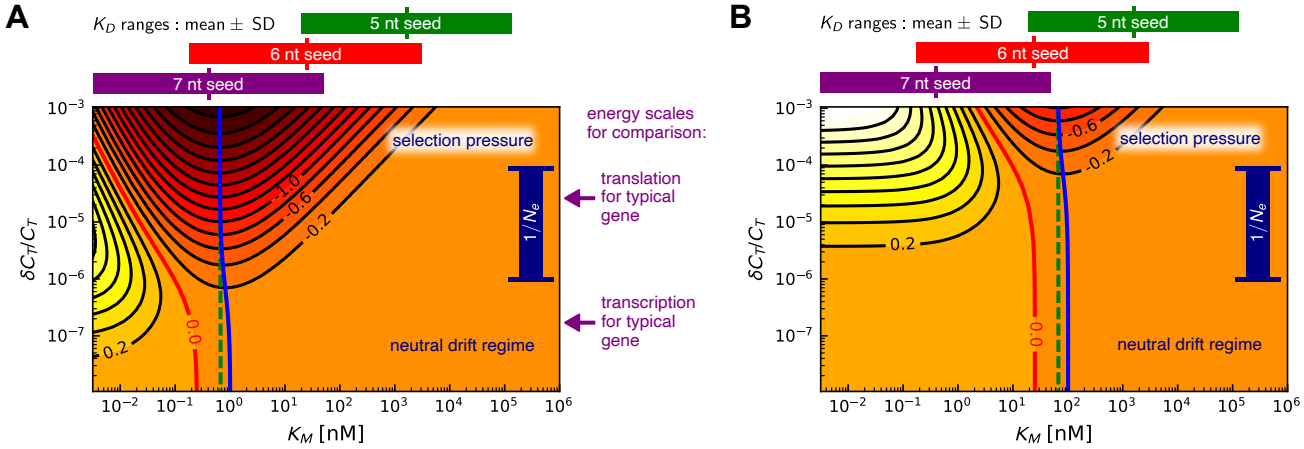

FIG. 2. Contour diagrams of noise filter error, $\log_{10} E$, in terms of Michaelis-Menten constant $K_{\rm M}$ and fractional metabolic cost $\delta C_T/C_T$ for a fixed protein output level $\bar{p}$ in the (**A**) single target and (**B**) 100 target cases. The spacing between contour values is 0.2. The minimum contour line for a given $\delta C_T/C_T$ corresponds to the most energetically efficient noise reduction, and this is achieved by $K_{\rm M}^*$ values along the blue curves. Similarly, the green dashed curves show the most efficient $K_{\rm M}^{\rm wk}$ values predicted by the WK optimal filtering theory. The red line is the boundary of the "dud" region on the left, inside which the miRNA regulation adds noise ($E > 1$) rather than mitigating it. To make a connection to physiological values of $K_{\rm M}$, we plot estimated K_D ranges for known miRNA-mRNA interactions above the plots for 5, 6, and 7 nt seed lengths. $K_{\rm D}$ sets a lower bound on $K_{\rm M}$ from Eq. [1]. Altogether, this shows that for a single typical target gene, the most economical noise reduction is likely to occur for 7 nt seeds, while 6 nt seeds become favorable for miRNAs with many targets. To better illustrate the evolutionary relevance, we also plot the dark blue bar showing inverse effective population sizes $N_e^{-1} = 10^{-6} - 10^{-4}$ typical for metazoans. As the fitness disadvantage due to metabolic costs become significant, $\delta C_T/C_T \gtrsim N_e^{-1}$, there is selective pressure on the organism driving it towards the $K_{\rm M}^*$ value. For comparison, we also show typical translation and transcription energy scales, based on Ref. [18].

single-target case. Interestingly, the $K_{\rm M}^*$ values in the single and multi-target case (about 0.65 nM and 65 nM respectively at high $\delta C_T/C_T$), are comparable to $K_{\rm M}$ measured in a fruit fly RNA interference pathway [19]. In the experiments, $K_{\rm M} = 1 \pm 0.2$ nM was found for a fully complementary interaction with a 7 nt seed, while single nucleotide mismatches in the seed binding (which in principle allow for a larger range of possible targets) boosted $K_{\rm M}$ by up to a factor of 82. While we do not yet have extensive experimental surveys of $K_{\rm M}$ values in miRNA (or related siRNA [small interfering RNA]) systems, it would be intriguing to check whether $K_{\rm M}$ tends to scale with the target population, as predicted by the optimal theory.

We can make the connection between the number of complementary matches (or seed length) and $K_{\rm M}$ more explicit. In SI Sec. S.IV we used the ViennaRNA server [22] to predict the ΔG^0 values of $\approx 10^4$ human miRNA seed sequences of length 7 nt, resulting in a distribution of $K_{\rm D}$ values which covered around ten decades on a logarithmic scale. The mean and standard deviation of $\log_{10}(K_{\rm D}/[1M]) = -9.5 \pm 2.2$ is shown as a purple bar above the plots in Fig. 2A,B. To mimic shorter seeds, we deleted 1 or 2 nucleotides from the sequence, to give the ranges $\log_{10}(K_{\rm D}/[1M]) = -7.6 \pm 2.0$ (6 nt, red bar) and -5.8 ± 1.9 (5 nt, green bar). As validation, the calculation was able to correctly reproduce measured $K_{\rm D}$ ranges for fully matched 7 nt seeds in fruit fly and mouse siRNA-target complexes. Since $K_{\rm D}$ sets the floor for $K_{\rm M}$, we see from Fig. 2A that the optimal $K_M^* \approx 1$ nM for the single-target case is unlikely to be accessible for 5 nt seeds. It becomes plausible for 6 nt seeds, and even more so for 7 nt seeds. Since $K_{\rm M}$ can be up to two decades larger than $K_{\rm D}$ [19], it is notable that 7 nt seeds densely cover the range of $K_{\rm D}$ values one or two decades smaller than $K_{\rm M}^*$. Extrapolating this pattern, longer seeds (≥ 8 nt) will be less likely than 7 nt to achieve $K_{\rm M}^*$. For the 100 target case (Fig. 2B) we see that the ideal seed length is shifted to 6 nt as a result of the increase in $K_{\rm M}^*$. Seeds of length 6 nt constitute 67% of a dataset of human and mouse miRNA seed sequences, with 7 nt seeds forming another 23% [23]. The preponderance of 6 nt seeds is in line with expectations if $K_{\rm M}$ was optimized, particularly since miRNAs will typically have many different targets.

Noise reduction can approach optimal linear filter performance. The filter analogy in Eq. [2] allows us to make an interesting comparison. For a given signal $s(t)$ and $n(t)$, we know that there is a filter function $H_{\rm wk}(t)$, the Wiener-Kolmogorov (WK) solution [24–26], which gives the best performance (smallest E) of all possible functions $H(t)$ for this type of linear noise filtering system (SI Sec. S.II B). We denote the corresponding value of error, $E_{\rm wk}^*$, and it serves as the overall lower bound on E. In our system, E reaches a minimum E^* at K_M^* for a given energetic cost, but is this minimum E^*

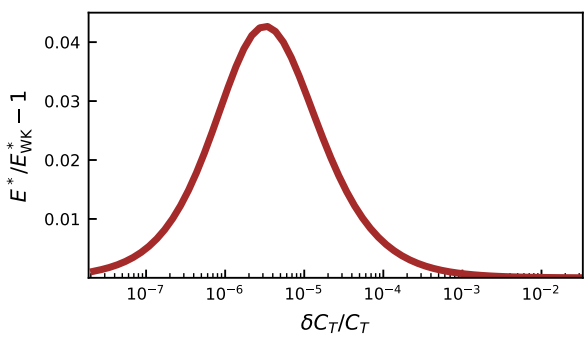

FIG. 3. The discrepancy between E^* and $E^*_{\rm wk}$, the error of the actual system and the WK optimal system respectively, at the most energetically efficient $K_{\rm M}$ for the single target case (Fig. 2A), at different values of the fractional cost $\delta C_T/C_T$. $E^*/E^*_{\rm wk} - 1 < 0.045$, so the miRNA system exhibits a performance close to WK optimality for this network motif.

comparable to $E^*_{\rm wk}$? In other words, can miRNA noise regulation approach an optimal WK filter? This type of comparison has recently proven fruitful in a variety of biological contexts [17, 27–33], for example yielding tight bounds on the fidelity of information transmission in signaling networks.

The miRNA system does not exactly realize true WK optimality, because the optimal filter function $H_{\rm wk}(t)$ cannot be precisely implemented by the miRNA regulation network. However, the affinity $K_{\rm M}^{\rm wk}$ predicted to be optimal by the WK theory for a given $\delta C_T/C_T$ (green dashed curve in Figs. 2A,B) is extremely close to $K_{\rm M}^*$ (blue curve). Fig. 3 shows the difference between the respective errors E^* and $E^*_{\rm wk}$ along these curves, as a function of $\delta C_T/C_T$. While $E^*/E^*_{\rm wk} - 1 > 0$, as expected, the difference is always smaller than 0.045, peaking at moderate cost values. As the cost $\delta C_T/C_T$ increases, E^* converges to $E^*_{\rm wk}$, and $K_{\rm M}^*$ similarly converges to $K_{\rm M}^{\rm wk}$. Notably, despite its constraints, the miRNA system can get quite close to WK optimal performance.

Evolutionary pressure on miRNA noise regulation. The previous two sections have argued that optimality in noise reduction (in the broader WK sense) is in principle approachable, and 6-7 nt seeds put miRNA systems within reach of achieving it. The final question we would like to consider is whether there would be any pressure from natural selection actually driving miRNA regulation toward optimality. From a population genetics perspective, let us say that the fitness of an organism with a particular miRNA regulatory network is f_R, while the same organism missing the network has fitness f_0. The selection coefficient $s = f_R/f_0 - 1$, quantifying the relative fitness, can be decomposed into two contributions, $s = s_a + s_c$ [18]. Here s_a is the adaptive advantage due to the regulation, for example resulting from protein noise reduction. The remaining part, $s_c < 0$, comes from the added metabolic cost of implementing the regulation. In order for s to be overall positive (and hence $f_R > f_0$) we would need an $s_a > 0$ that is larger in magnitude than the cost, $s_a > |s_c|$.

The evolution of miRNA systems, however, poses a conundrum: a newly arisen miRNA, before any selective fine-tuning of its seed sequence, could potentially target hundreds of mRNA in a random manner, making $s_a < 0$ due to deleterious effects on existing genetic networks. So how would advantageous miRNA regulation eventually emerge? The solution to this problem, as argued by Chen and Rajewsky [2], is for new miRNA systems to be expressed at very low levels, such that s_a, even if negative, has a negligible magnitude. In this regime however $s \approx s_c < 0$, which is still an overall fitness disadvantage. Would such a deleterious variant survive in a population long enough for further mutations to confer a significant positive s_a? The answer depends on the magnitude of $|s_c|$ relative to a threshold known as the "drift barrier" [34]. This threshold is set by N_e^{-1}, where N_e is a measure of genetic diversity known as the effective population size [35]. N_e is the size of an idealized population that shows the same changes in genetic diversity per generation (due to genetic drift, or random sampling of variants) as the actual population. N_e is generally smaller than the real population size, and among more complex eukaryotes (like vertebrates) can be as small as $10^4 - 10^6$ [18, 36]. If $|s_c| \ll N_e^{-1}$, selection against the variant is weak, even when $s_c < 0$, so it can survive in a population via genetic drift as an effectively neutral mutant and even eventually take over (with a fixation probability roughly given by its initial fraction in the population [37]). On the other hand, if $|s_c| \gg N_e^{-1}$, then the costs are sufficiently high that selection would efficiently weed out the variant from the population, unless there were compensating advantages, $s_a \gtrsim |s_c|$.

A recent derivation based on a general bioenergetic growth model for organisms allows us to make the above discussion more quantitative: it showed that $s_c \approx -\ln(R_b)\delta C_T/C_T$, where R_b is the mean number of offspring per individual [38]. Assuming that $\ln(R_b)$ does not change the order of magnitude, we can thus use the fractional cost $\delta C_T/C_T$ as a proxy for $|s_c|$, and compare it to N_e^{-1}. The blue range bars on the right in Fig. 2A,B show possible N_e^{-1} values 10^{-6}–10^{-4} for higher-order eukaryotes, separating a selection pressure regime at large $\delta C_T/C_T$ from a neutral drift regime at low $\delta C_T/C_T$. For comparison we also show the cost scales for transcription and translation of a typical gene [18], which indicate that transcription is not generally not under selective pressure in these organisms, while translation may be.

At the lowest expression levels, a newly evolved miRNA system could survive in the neutral drift regime, even with a non-optimal $K_{\rm M}$. There is limited noise reduction achievable in this regime, since the contours indicating E significantly smaller than 1 require larger $\delta C_T/C_T$, particularly for the multi-target case (Fig. 2B). Thus the initial evolution could be imagined as a random walk near the bottom edge of the diagram. Mutations

that led to greater expression of an miRNA, moving up on the cost scale, would hit against the drift barrier, and would more likely survive if they came with compensating fitness advantages. In a context where noise control was beneficial, this would mean being funneled up the region where $K_{\rm M}$ is close to $K^*_{\rm M}$, the path that confers the largest noise reduction as expression levels rise. Once the costs are above the drift barrier, there would be significant selective pressure to optimize $K_{\rm M}$. In this high cost regime there is a direct tradeoff between extra metabolic expenditure and noise filtering, with $\Delta M_\nu \sim M_\nu/E$, and hence $s_c \sim -M_\nu/E$. Our theory predicts that any compensating fitness advantage $s_a > 0$ would have to grow like E^{-1} or faster as E decreases, in order for the regulation to be viable in the long term.

Inferring closeness to optimality in experimental systems. Given the achievability of optimal affinities $K^*_{\rm M}$ for noise filtering (based on seed lengths), and the evolutionary pressures that could drive a system toward this optimality, is there any corroborating experimental evidence? While we do not have simultaneous measurements yet of $K_{\rm M}$ and noise regulation in specific systems, there is a way of approximately inferring the ratio $K_{\rm M}/K^*_{\rm M}$ from fitting our model to existing data on miRNA noise suppression. The two experiments we focus on are assays that take the 3'UTR regions of endogenous mRNAs and combine them with fluorescent reporters like *mCherry*—in one case the 3'UTR was from the *Lats2* gene in mouse embryonic stem cells [5], and in the other from the *sens* gene in *Drosophila* wing disc cells [39]. These 3'UTR regions have binding sites for endogenous miRNA, and the assay demonstrates that miRNA interaction suppresses protein noise for these genes, by quantifying reporter protein fluctuations in the wildtype compared to a mutant system where the binding sites were altered, inhibiting miRNA regulation. In both cases the noise suppression has potential functional roles—Lats2 is involved in regulating the cell cycle, apoptosis, and differentiation [40, 41], and excess noise in Sens protein levels leads to disordered sensory patterning in wing disc cells [39].

Thus there is reason to suspect that there could have been selective pressure on the mRNA-miRNA affinity for these genes. To investigate this hypothesis, we took the available experimental data from the two studies and extracted the error E and regulation strength R as a function of protein expression, since there was a distribution of values for $\bar{p}$ (or reporter intensity) $\propto \bar{m}$ among the population of cells in each system. (The full details of the data analysis can be found in SI Sec. S.V). For a given cell, we know that $K^*_{\rm M}$ is proportional to the mRNA concentration $\bar{m}$ from Eq. [4], and thus should be higher in high-expression versus low-expression cells. On the other hand, the types of miRNA and binding sites are the same between cells, so the actual affinity $K_{\rm M}$ should also be the same in each cell. As before, we use a simple version of the multi-miRNA, multi-target theory assuming similar

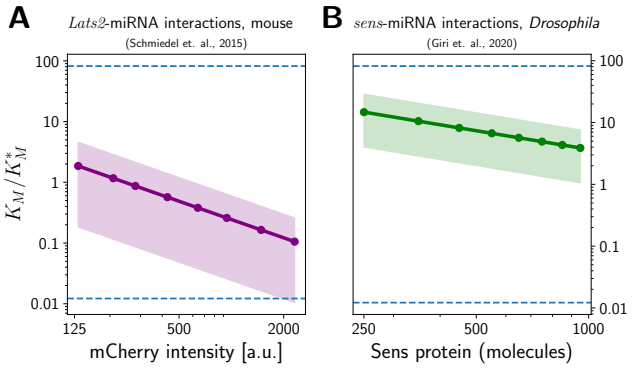

FIG. 4. Analysis of optimality for miRNA-mRNA affinity in data from two experimental systems. We show the inferred ratio of actual to optimal Michaelis-Menten constants $K_{\rm M}/K^*_{\rm M}$, as a function of protein expression level in individual cells, for miRNA regulation of: (A) the *Lats2* gene in mouse embryonic stem cells [5]; (B) the *sens* gene in *Drosophila* wing disc cells [39]. The points connected by a line are theoretical best-fits using typical parameter values from SI Table S1, while the colored regions represent uncertainties due to varying the system parameters γ_μ and σ_μ over biologically plausible ranges. The horizontal axis reflects different quantification of protein expression in the two experiments, either in fluorescent reporter intensity (Lats2) or copy number (Sens). The dashed lines represent the maximum positive or negative fold-change in $K_{\rm M}$ observed experimentally from single nucleotide differences in seed matching in Ref. [19].

biochemical parameters for each miRNA-mRNA interaction, so we can interpret $K_{\rm M}$ to be an average affinity across the ensemble of miRNAs interacting with the targets. To summarize, $K_{\rm M}$ is fixed but $K^*_{\rm M}$ varies between cells, and we can ask whether the ratio $K_{\rm M}/K^*_{\rm M}$ is close to 1 (optimality) over the cell population. Fig. 4 shows the results for this ratio as a function of protein expression in the *Lats2* and *sens* systems, derived from fitting our model to the data. The points connected via lines correspond to fits using the typical parameter values in SI Table S1, while the surrounding colored regions represent uncertainties due to imperfect knowledge of the parameters γ_μ and σ_ϵ (see SI Sec. S.V). Despite this uncertainty, the inferred $K_{\rm M}/K^*_{\rm M}$ ratio is roughly within an order of magnitude of 1, and for the *Lats2* system even crosses 1. The horizontal dashed lines above and below 1 indicate a factor of 82 in either direction, representing the maximum magnitude of fold change in $K_{\rm M}$ observed from a single nucleotide difference in seed matching in another experiment [19]. This range emphasizes the relative narrowness of the observed $K_{\rm M}/K^*_{\rm M}$ ratio, with the affinity fine-tuned to within a nucleotide of optimality. Moreover, this is true across the whole span of protein expression observed in the experiments, which is also physiologically relevant for these cells. So even though low-expression cells and high-expression cells have different optimal parameters (and error bounds E^*), they all are fairly close to their respective optima. While there is still work to

be done in narrowing down parameter uncertainties and extending the analysis to other systems in future studies, this initial analysis is consistent with evolutionary pressures shaping miRNA-mRNA affinity in systems where noise reduction is important.

Discussion and conclusions. Putting all the elements of our theory together, we show that if noise control confers a fitness advantage for a particular miRNA system, there is selective pressure driving it toward an optimal miRNA-mRNA affinity, as described by the Michaelis-Menten constant K_{M}. Remarkably, the optimal value K_{M}^{*} in Eq. [4] can be achieved by a fairly narrow range of seed lengths (6-7 nt), which happen to make up the vast majority of miRNA seeds. While we argue the plausibility of this key result based on realistic ranges of biological parameters, Eq. [4] opens the way for future experimental validation in specific, fully characterized systems. Such validation should be practical, since the equation only involves a small number of biochemical parameters. The noise reduction at optimality approaches the performance of a Wiener-Kolmogorov filter, the best possible linear noise filter. If true, such optimization would be a striking example of metabolic costs directly shaping the course of evolution for a biochemical network in eukaryotes. This is unusual in itself, because eukaryotes are generally less likely to prioritize energy efficiency relative to prokaryotes, which have higher effective population sizes and thus lower drift barriers [18]. Beyond the implications for miRNA evolution, the theory could also find applications in the design of synthetic circuits with 3'UTR engineering and artificial miRNAs [42–45].

While the simple theory of miRNA-mRNA interaction used here is sufficient to describe certain experiments [5], there are a variety of model assumptions that can be relaxed in future investigations, to test the conclusions more broadly. For example, one aspect was ignored in the current model: many miRNAs can bind to multiple sites on a single target, with potentially different affinities, as well as exhibit varied affinities more generally across multiple mRNA targets. While we do not expect this heterogeneity to qualitatively change the overall results, it will be interesting to see how it shifts the relations between affinity, metabolic costs, and noise reduction. Multiple binding sites / targets can also lead to nonlinear phenomena like ultrasensitivity and bistability in miRNA-mRNA systems [46], which in turn could give noise reduction additional functional implications, like inhibiting stochastic switching between different cellular states. More complex models can also consider the role of miRNA in mitigating the effects of (possibly nonstationary) environmental fluctuations, along with the noise due to cellular processes [8, 47].

Though our work focuses on noise control via miRNA regulation, it is also important to keep in mind that noise control can be implemented via other mechanisms, and that miRNA themselves have other functional roles. Protein noise can be reduced (maintaining the same expression level) by increasing transcription and either decreasing translation rates or enhancing mRNA degradation [48, 49], and this would avoid noise added due to miRNA interactions [50]. Degradation rates could for example be fine-tuned by the length of mRNA poly(A) tails [51]. However this mechanism is non-specific, while miRNAs allow for selective noise control via seed recognition. Though our results suggest a formative role for noise reduction in shaping miRNA-mRNA affinity, the filtering capacity likely co-evolved with other miRNA functions such gene silencing and crosstalk [52]. Perhaps in some cases noise regulation is simply a complementary benefit of a far more complex utility scheme.

Take for instance the scenario where there is selective pressure for gene silencing via miRNA, in other words suppressing the protein level $\bar{p}$ for a particular gene. Naively, one might expect this pressure to always favor smaller values of K_{M}, since higher affinities can provide the same amount of suppression using fewer miRNA (and hence at smaller metabolic cost). But as we saw in our theoretical analysis (i.e. the "dud" regime on the left in the panels of Fig. 2), smaller K_{M} also introduces extra noise into the system. If this noise becomes sufficiently high that it has deleterious effects, like unwanted stochastic transitions between cellular states, then there will be a countervailing pressure to increase K_{M}. The balance between these two effects might result in a metabolic sweet spot for affinity, analogous to the one we described in our model, though the posited functional role of the miRNA system was different.

Viral miRNAs present another example of alternative functional roles. These miRNAs exploit the host metabolism, and are likely useful for non-noise-related tasks like evading the host immune response [53]. Notably, though viruses lack their own metabolic machinery, there can still be selective pressures on viral miRNA expression as part of the overall energetic costs associated with viral copying [54]. Ultimately a deeper understanding of miRNA evolution will require larger-scale models of its full regulatory context, coupled with *in vivo* experiments to explore the tangled effects of function, metabolic costs, and fitness.

ACKNOWLEDGMENTS

We thank Amit Singh Vishen for helpful discussions and Xingbo Yang for a critical reading of the manuscript and insightful comments.

Data availability

Code related to the calculations, routines for plotting the figures, and all the associated datasets are available at the following Github repository: `https://github.com/hincz-lab/microrna/`.

[1] X. Yang, M. Heinemann, J. Howard, G. Huber, S. Iyer-Biswas, G. Le Treut, M. Lynch, K. L. Montooth, D. J. Needleman, S. Pigolotti, *et al.*, Physical bioenergetics: Energy fluxes, budgets, and constraints in cells, Proc. Natl. Acad. Sci. **118**, e2026786118 (2021).

[2] K. Chen and N. Rajewsky, The evolution of gene regulation by transcription factors and micrornas, Nat. Rev. Genet. **8**, 93 (2007).

[3] D. P. Bartel, Metazoan micrornas, Cell **173**, 20 (2018).

[4] M. Osella, C. Bosia, D. Corá, and M. Caselle, The role of incoherent microrna-mediated feedforward loops in noise buffering, PLoS Comput. Biol. **7**, e1001101 (2011).

[5] J. M. Schmiedel, S. L. Klemm, Y. Zheng, A. Sahay, N. Blüthgen, D. S. Marks, and A. van Oudenaarden, Microrna control of protein expression noise, Science **348**, 128 (2015).

[6] V. Siciliano, I. Garzilli, C. Fracassi, S. Criscuolo, S. Ventre, and D. Di Bernardo, Mirnas confer phenotypic robustness to gene networks by suppressing biological noise, Nat. Comm. **4**, 2364 (2013).

[7] E. Hornstein and N. Shomron, Canalization of development by micrornas, Nat. Genet. **38**, S20 (2006).

[8] M. S. Ebert and P. A. Sharp, Roles for micrornas in conferring robustness to biological processes, Cell **149**, 515 (2012).

[9] K. J. Peterson, M. R. Dietrich, and M. A. McPeek, Micrornas and metazoan macroevolution: insights into canalization, complexity, and the cambrian explosion, Bioessays **31**, 736 (2009).

[10] C. Alberti and L. Cochella, A framework for understanding the roles of mirnas in animal development, Development **144**, 2548 (2017).

[11] D. T. Gillespie, The chemical langevin equation, J. Chem. Phys. **113**, 297 (2000).

[12] G. Tkačik and A. M. Walczak, Information transmission in genetic regulatory networks: a review, Journal of Physics: Condensed Matter **23**, 153102 (2011).

[13] J. Tsang, J. Zhu, and A. Van Oudenaarden, Microrna-mediated feedback and feedforward loops are recurrent network motifs in mammals, Mol. Cell **26**, 753 (2007).

[14] N. J. Martinez and A. J. Walhout, The interplay between transcription factors and micrornas in genome-scale regulatory networks, Bioessays **31**, 435 (2009).

[15] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel, The impact of micrornas on protein output, Nature **455**, 64 (2008).

[16] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky, Widespread changes in protein synthesis induced by micrornas, Nature **455**, 58 (2008).

[17] D. Hathcock, J. Sheehy, C. Weisenberger, E. Ilker, and M. Hinczewski, Noise filtering and prediction in biological signaling networks, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 16 (2016).

[18] M. Lynch and G. K. Marinov, The bioenergetic costs of a gene, Proc. Natl. Acad. Sci. **112**, 15690 (2015).

[19] L. M. Wee, C. F. Flores-Jasso, W. E. Salomon, and P. D. Zamore, Argonaute divides its rna guide into domains with distinct functions and rna-binding properties, Cell **151**, 1055 (2012).

[20] M. Del Giudice, S. Bo, S. Grigolon, and C. Bosia, On the role of extrinsic noise in microrna-mediated bimodal gene expression, PLoS Comput. Biol. **14**, e1006063 (2018).

[21] E. Ferro, C. E. Bena, S. Grigolon, and C. Bosia, microrna-mediated noise processing in cells: A fight or a game?, Comput. Struct. Biotechnol. J. **18**, 642 (2020).

[22] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, Viennarna package 2.0, Algorithms Mol. Biol. **6**, 1 (2011).

[23] D. C. Ellwanger, F. A. Büttner, H.-W. Mewes, and V. Stümpflen, The sufficient minimal set of mirna seed types, Bioinformatics **27**, 1346 (2011).

[24] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*, Vol. 2 (MIT press Cambridge, 1949).

[25] A. N. Kolmogorov, Interpolation and extrapolation of stationary random sequences, Izv. Akad. Nauk SSSR Ser. Mat. **5**, 3 (1941).

[26] H. W. Bode and C. E. Shannon, A simplified derivation of linear least square smoothing and prediction theory, Proc. Inst. Radio. Engin. **38**, 417 (1950).

[27] M. Hinczewski and D. Thirumalai, Cellular signaling networks function as generalized wiener-kolmogorov filters to suppress noise, Phys. Rev. X **4**, 041017 (2014).

[28] N. B. Becker, A. Mugler, and P. R. ten Wolde, Optimal prediction by cellular signaling networks, Phys. Rev. Lett. **115**, 258103 (2015).

[29] M. Hinczewski and D. Thirumalai, Noise control in gene regulatory networks with negative feedback, J. Phys. Chem. B (2016).

[30] D. Hartich and U. Seifert, Optimal inference strategies and their implications for the linear noise approximation, Phys. Rev. E **94**, 042416 (2016).

[31] P. R. ten Wolde, N. B. Becker, T. E. Ouldridge, and A. Mugler, Fundamental limits to cellular sensing, J. Stat. Phys. **162**, 1395 (2016).

[32] C. Zechner, G. Seelig, M. Rullan, and M. Khammash, Molecular circuits for dynamic noise filtering, Proc. Natl. Acad. Sci. U. S. A. **113**, 4729 (2016).

[33] T.-L. Wang, B. Kuznets-Speck, J. Broderick, and M. Hinczewski, The price of a bit: energetic costs and the evolution of cellular signaling, bioRxiv , 2020.10.06.327700 (2020).

[34] W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch, Drift-barrier hypothesis and mutation-rate evolution, Proc. Natl. Acad. Sci. **109**, 18488 (2012).

[35] J. H. Gillespie, *Population Genetics: A Concise Guide* (JHU Press, 2010).

[36] B. Charlesworth, Effective population size and patterns of molecular evolution and variation, Nat. Rev. Genet. **10**, 195 (2009).

[37] M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).

[38] E. Ilker and M. Hinczewski, Modeling the growth of organisms validates a general relation between metabolic costs and natural selection, Phys. Rev. Lett. **122**, 238101 (2019).

[39] R. Giri, D. K. Papadopoulos, D. M. Posadas, H. K. Potluri, P. Tomancak, M. Mani, and R. W. Carthew, Ordered patterning of the sensory system is susceptible to stochastic features of gene expression, eLife **9**, e53638

(2020).

[40] N. Yabuta, N. Okada, A. Ito, T. Hosomi, S. Nishihara, Y. Sasayama, A. Fujimori, D. Okuzaki, H. Zhao, M. Ikawa, *et al.*, Lats2 is an essential mitotic regulator required for the coordination of cell division, Journal of Biological Chemistry **282**, 19259 (2007).

[41] N. Furth and Y. Aylon, The lats1 and lats2 tumor suppressors: beyond the hippo pathway, Cell Death & Differentiation **24**, 1488 (2017).

[42] F. J. Navarro and D. C. Baulcombe, mirna-mediated regulation of synthetic gene circuits in the green alga chlamydomonas reinhardtii, ACS Synth. Biol. **8**, 358 (2019).

[43] A. Kotowska-Zimmer, M. Pewinska, and M. Olejniczak, Artificial mirnas as therapeutic tools: Challenges and opportunities, Wiley Interdiscip. Rev. RNA **12**, e1640 (2021).

[44] T. Kang, T. Quarton, C. M. Nowak, K. Ehrhardt, A. Singh, Y. Li, and L. Bleris, Robust filtering and noise suppression in intragenic mirna-mediated host regulation, iScience **23** (2020).

[45] L. Wei, S. Li, P. Zhang, T. Hu, M. Q. Zhang, Z. Xie, and X. Wang, Characterizing microrna-mediated modulation of gene expression noise and its effect on synthetic gene circuits, Cell Rep. **36** (2021).

[46] X.-J. Tian, H. Zhang, J. Zhang, and J. Xing, Reciprocal regulation between mrna and microrna enables a bistable switch that directs cell fate decisions, FEBS Lett. **590**, 3443 (2016).

[47] X. Li, J. J. Cassidy, C. A. Reinke, S. Fischboeck, and R. W. Carthew, A microrna imparts robustness against environmental fluctuation during development, Cell **137**, 273 (2009).

[48] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. Van Oudenaarden, Regulation of noise in the expression of a single gene, Nat. Genet. **31**, 69 (2002).

[49] A. Raj and A. van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its consequences, Cell **135**, 216 (2008).

[50] R. Fan and A. Hilfinger, The effect of microrna on protein variability and gene expression fidelity, Biophys. J. **122** (2023).

[51] O. S. Rissland, A. O. Subtelny, M. Wang, A. Lugowski, B. Nicholson, J. D. Laver, S. S. Sidhu, C. A. Smibert, H. D. Lipshitz, and D. P. Bartel, The influence of micrornas and poly (a) tail length on endogenous mrna–protein complexes, Genome Biol. **18**, 1 (2017).

[52] M. Jens and N. Rajewsky, Competition between target sites of regulators shapes post-transcriptional gene regulation, Nat. Rev. Genet. **16**, 113 (2015).

[53] R. L. Skalsky and B. R. Cullen, Viruses, micrornas, and host interactions, Annu. Rev. Microbiol. **64**, 123 (2010).

[54] G. Mahmoudabadi, R. Milo, and R. Phillips, Energetic cost of building a virus, Proc. Natl. Acad. Sci. **114**, E4324 (2017).

Supplemental Material: Bioenergetic costs and the evolution of noise regulation by microRNAs

Efe Ilker[1] and Michael Hinczewski[2]

[1] *Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*

[2] *Department of Physics, Case Western Reserve University, Cleveland OH 44106*

CONTENTS

S.I. Analysis of microRNA-mediated noise regulation 1
A. Chemical Langevin formalism of microRNA-mediated gene expression 1
1. General birth-death processes 1
2. Protein production 2
3. Steady-state properties, regulation strength, bound fraction of miRNAs: 4
B. Protein expression noise 4
C. Bioenergetic costs of miRNA regulation 5

S.II. The noise-filter formulation of the microRNA regulated system 6
A. Realizing miRNA regulation as a noise filtering mechanism 6
B. Wiener-Kolmogorov optimality of protein output noise level 7

S.III. Calculation of phase diagrams 9
A. Contour plots, fitness costs, and Michaelis-Menten constant 9
B. Minimal noise for the same metabolic cost 10

S.IV. Estimating dissociation constants K_{D} for seed sequences of varying length 11

S.V. Analysis of experimental microRNA systems 11

References 16

S.I. ANALYSIS OF MICRORNA-MEDIATED NOISE REGULATION

A. Chemical Langevin formalism of microRNA-mediated gene expression

1. General birth-death processes

We start with a general description of chemical Langevin (CL) dynamics, which is the basis for our theoretical model. A biochemical reaction network consists of N different types of biomolecules with time-dependent populations $n_j(t)$, $j = 1, \ldots, N$. The rate of change of a given population, $dn_j(t)/dt$, can be expressed in terms of a number of possible stochastic reactions, which we will label with an index $\rho = 1, \ldots, r$, where r is the total number of

reactions. The CL dynamics of the birth-death process for biomolecule j can be written as [1]:

$$\frac{dn_j(t)}{dt} = \sum_{\rho=1}^{r} \Delta_{\rho j}\pi_\rho(t) + \sum_{\rho=1}^{r} \Delta_{\rho j}\sqrt{\pi_\rho(t)}\eta_\rho(t) \tag{S1}$$

where $\Delta_{\rho j}$ are the stoichiometric coefficients representing the change in the population of j in reaction ρ, $\pi_\rho(t)$ are the rates of reactions and $\eta_\alpha(t)$ are independent Gaussian noise terms at time t which satisfy $\langle \eta_\rho(t)\eta_\beta(t')\rangle = \delta_{\rho\beta}\delta(t-t')$ where $\delta_{\rho\beta}$ and $\delta(t-t')$ are Kronecker delta and Dirac delta functions, respectively. As the reactions for the production and degradation of j are uncorrelated, we can add up the uncorrelated noise terms, and obtain:

$$\frac{dn_j(t)}{dt} = \sum_{\rho=1}^{r} \Delta_{\rho j}\pi_\rho(t) + \sqrt{\alpha_j(t)}\eta_j(t) \tag{S2}$$

where $\alpha_j(t) = \sum_{\rho=1}^{r} \Delta_{\rho j}^2 \pi_\rho(t)$. Note that though the η_j's are uncorrelated at different times, there can be non-zero correlations between the dynamics of different populations at time t if they share the same reactions.

2. *Protein production*

To apply the CL framework to microRNA regulation of gene expression, we focus on the following biomolecular species and reactions, following the model of Ref. [2]:

$$\begin{aligned}
\emptyset &\underset{d_m}{\overset{\nu_R}{\rightleftarrows}} m \\
m &\xrightarrow{k_p} m + p \\
p &\xrightarrow{d_p} \emptyset \\
\emptyset &\underset{d_\mu}{\overset{\nu_\mu}{\rightleftarrows}} \mu \\
m + \mu &\underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} m_b \\
m_b &\xrightarrow{d_{m\mu}} \mu \\
m_b &\xrightarrow{d_\mu} m
\end{aligned}$$

Here m, p, μ, m_b represent respectively mRNA, protein, miRNA and the miRNA-mRNA complex. These reactions include birth-death processes of individual species as well as the formation and degradation of complexes. When the miRNA-mRNA complex m_b degrades, the miRNAs are recycled. For convenience, we assume a similar property when miRNAs spontaneously degrade. This model has been shown to agree well with experimental systems of miRNA-mediated gene expression and noise reduction [2, 3].

The reactions above give the following mass-action kinetics for the concentrations of the four species (denoted by brackets):

$$\frac{d[\mu]}{dt} = \nu_\mu - d_\mu[\mu] - k_{\text{on}}[m][\mu] + (k_{\text{off}} + d_{m\mu})[m_b] \tag{S3a}$$

$$\frac{d[m]}{dt} = \nu_R - d_m[m] - k_{on}[m][\mu] + (k_{\text{off}} + d_\mu)[m_b] \tag{S3b}$$

$$\frac{d[m_b]}{dt} = k_{\text{on}}[m][\mu] - (k_{\text{off}} + d_{m\mu} + d_\mu)[m_b] \tag{S3c}$$

$$\frac{d[p]}{dt} = k_p[m] - d_p[p]. \tag{S3d}$$

The total miRNA concentration is a sum of free and bound miRNAs, i.e., $[\mu_{tot}] = [\mu] + [m_b]$. Assuming the binding/unbinding rate of the mRNA-microRNA complex is much faster than other dynamics, we obtain a quasi-steady-state approximation leading to $[m_b] = \frac{[m][\mu_{tot}]}{K_{\text{M}}+[m]}$, with Michaelis-Menten constant $K_{\text{M}} = K_{\text{D}} + k_{\text{cat}}/k_{\text{on}}$ expressed in terms of the dissociation constant $K_{\text{D}} = k_{\text{off}}/k_{\text{on}}$ and an effective catalytic rate constant $k_{\text{cat}} = d_{m\mu} + d_\mu$ for the miRNA-catalyzed degradation reaction. With these simplifications, we obtain

$$\frac{d[\mu_{tot}]}{dt} = \nu_\mu - d_\mu[\mu_{tot}] \tag{S4a}$$

$$\frac{d[m]}{dt} = \nu_R - d_m[m] - \frac{d_{m\mu}}{\kappa_M + [m]}[\mu_{tot}][m] \tag{S4b}$$

$$\frac{d[p]}{dt} = k_p[m] - d_p[p]. \tag{S4c}$$

Note that the description in terms of chemical concentrations ignores the fact that the crowded cell interior is not an ideal solution, and that in principle the model can be made more accurate by replacing concentrations with chemical activities in all the reaction rate terms. However, the activity coefficients depend on the concentrations of all chemical species in the cytoplasm (including species not explicitly present in the model), which makes mathematical analysis intractable. The simplest approximation is to assume constant activity coefficients, based on the steady-state values of all the concentrations. In this scenario the activity coefficients can be absorbed into effective rate constants, and the structure of the kinetic equations above stays the same.

Converting concentrations to copy numbers (denoted without brackets) by multiplying with the cell volume V, and using Eqs. (S1)-(S2), we can now introduce the corresponding CL equations for the model:

$$\frac{d\mu_{tot}(t)}{dt} = \nu_\mu - d_\mu\mu_{tot}(t) + \sqrt{\alpha_\mu}\eta_\mu(t) \tag{S5a}$$

$$\frac{dm(t)}{dt} = \nu_R - d_m m(t) - \frac{d_{m\mu}}{\kappa_M + m(t)}\mu_{tot}(t)m(t) + \sqrt{\alpha_m}\eta_m(t) + \sqrt{\alpha_b}\eta_b(t) \tag{S5b}$$

$$\frac{dp(t)}{dt} = k_p m(t) - d_p p(t) + \sqrt{\alpha_p}\eta_p(t), \tag{S5c}$$

where $\langle \eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t-t')$. We will provide a detailed derivation of the noise amplitudes $\alpha_j(t)$ in Section S.I B.

3. *Steady-state properties, regulation strength, bound fraction of miRNAs:*

At the steady-state, the mean populations for our dynamical system are $\bar{p} = \frac{k_p}{d_p}\bar{m}$, $\bar{\mu}_{\rm tot} = \nu_\mu/d_\mu$, $\bar{m} = \frac{\nu_R}{d_m}(1-R)$, where the regulation strength R is defined as:

$$R = \frac{\theta_\mu \bar{\mu}_{\rm tot}}{\theta_\mu \bar{\mu}_{\rm tot} + \gamma_m \bar{m}}. \tag{S6}$$

Here $\gamma_m \equiv d_m/d_{m\mu}$ and θ_μ is the bound fraction of miRNAs,

$$\theta_\mu \equiv \frac{\bar{m}_b}{\bar{\mu}_{\rm tot}} = \frac{\bar{m}}{K_{\rm M} + \bar{m}}. \tag{S7}$$

In order to compare protein noise at different levels of miRNA regulation (including the no regulation case), we fix the mean protein population by assuming the mRNA production rate ν_R changes with R to compensate for the effects of regulation,

$$\nu_R = \frac{\nu_0}{1-R} = \frac{d_m \bar{m}}{1-R} \tag{S8}$$

where $\nu_0 = \nu_R(R=0)$.

B. Protein expression noise

The fluctuating molecular populations are defined in the main text as $x(t) = \bar{x} + \delta x(t)$ where $\delta x(t)$ is the variation from the mean level $\bar{x}$ and the species of interest are $x(t) \equiv \{\mu_{tot}(t), m(t), p(t)\}$. Since we focus on the stationary-state fluctuations, we may treat the noise amplitudes α_j as time-independent. We can write these in terms of the equilibrium populations,

$$\alpha_\mu = 2d_\mu \bar{\mu}_{\rm tot},\ \alpha_m = 2d_m \bar{m},\ \alpha_b = \alpha_m R/(1-R),\ \alpha_p = 2d_p \bar{p}, \tag{S9}$$

where $\bar{\mu}_{\rm tot}$, $\bar{m}$, and $\bar{p}$ are mean steady-state population levels for miRNA, mRNA and protein respectively. Using the linear-noise approximation in Eq. (S5) and taking the Fourier transform leads to

$$\delta\mu(\omega) = \frac{\sqrt{\alpha_\mu}\eta_\mu(\omega)}{d_\mu - i\omega} \tag{S10a}$$

$$\delta m_R(\omega) = \frac{\sqrt{\alpha_m}\eta_m(\omega) + \sqrt{\alpha_m R/(1-R)}\eta_m(\omega) - d_{m\mu}\theta_\mu \delta\mu(\omega)}{c_m - i\omega} \tag{S10b}$$

$$\delta p_R(\omega) = \frac{k_p \delta m_R(\omega)}{d_p - i\omega} + \frac{\sqrt{\alpha_p}\eta_p(\omega)}{d_p - i\omega}, \tag{S10c}$$

with $c_m = d_m(1-R\theta_\mu)/(1-R)$. In Fourier space, the noise correlations satisfy $\langle \eta_i(\omega)\eta_j(\omega')\rangle = 2\pi\delta_{ij}\delta(\omega+\omega')$. The limit $R \to 0$ corresponds to an unregulated system, with $c_m \to d_m$ and $d_{m\mu}\theta_\mu\delta\mu(\omega) \to 0$ in Eq. (S10b).

The main measure that we use to assess noise levels in protein populations is the Fano factor $F \equiv \sigma_p^2/\bar{p}$ where the variance $\sigma_p^2 \equiv \langle(\delta p(t))^2\rangle = \langle(\delta p(0))^2\rangle = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\langle(\delta p_R(\omega)\delta p_R(-\omega))\rangle$. Using Eq. (S10c) and the correlation properties of the noise terms $\eta_i(\omega)$ in Fourier space, we can calculate the Fano factors F_0 and F_R for the unregulated and regulated systems respectively:

$$F_0 = 1 + \frac{k_p}{d_m + d_p}, \tag{S11a}$$

$$F_R = 1 + \frac{k_p}{(c_m + d_p)(1 - R\theta_\mu)} + \frac{d_{m\mu}\theta_\mu(c_m + d_\mu + d_p)k_p R}{(c_m + d_\mu)(c_m + d_p)(d_\mu + dp)(1 - R\theta_\mu)}. \tag{S11b}$$

The third term in F_R is a result of coupling with microRNA pool noise (extrinsic noise). We can quantify the relative protein noise levels in an miRNA-regulated system versus an unregulated one with the quantity

$$E = \frac{F_R - 1}{F_0 - 1}. \tag{S12}$$

To simplify our analysis, we define the following nondimensional parameters:

$$\gamma_m \equiv \frac{d_m}{d_{m\mu}}, \;\; \gamma_\mu \equiv \frac{d_\mu}{d_{m\mu}}, \;\; \phi \equiv \frac{d_p}{d_m}. \tag{S13}$$

Assuming the degradation rate constant of proteins is slower than that of RNAs, i.e. $\phi \ll 1$, we can derive an approximate expression for E,

$$E \approx \frac{(1 - R)(\gamma_\mu + R\theta_\mu)}{\gamma_\mu(1 - R\theta_\mu)^2}. \tag{S14}$$

C. Bioenergetic costs of miRNA regulation

To estimate bioenergetic costs, we follow the Lynch & Marinov analysis [4], where the transcriptional cost of a typical gene in eukaryotic cells is given by $C_T \equiv c_g + M_\nu t_r$. Here c_g are costs due to synthesis (growth), M_ν is the energy consumption rate for maintenance, and t_r is the lifetime of the cell. For long enough cell-division times, the maintenance term dominates, and hence $C_T \approx M_\nu$. For the unregulated system, this amounts to $M_\nu = \nu_0\epsilon_m$, where ϵ_m is the maintenance energy cost per mRNA (i.e. the turnover cost of recycling its nucleotides after degradation into new mRNA). By contrast, the cost for the miRNA regulated system consists of two contributions: increased transcription of mRNAs in order to maintain $\bar{p}$ and transcription of miRNAs. The additional cost relative to the unregulated case, ΔM_ν, can be expressed as $\Delta M_\nu = (\nu_R - \nu_0)\epsilon_m + \nu_\mu\epsilon_\mu$, where ϵ_μ is the maintenance cost per miRNA. By substituting $\nu_R = \nu_0/(1 - R)$, $\nu_\mu = d_\mu\bar{\mu}_{\text{tot}}$, $\nu_0 = d_m\bar{m}$ from the definitions in Section S.I A 3, and inverting Eq. (S6) for $\bar{\mu}_{\text{tot}}$, we obtain Eq. (3) in the main text,

$$\Delta M_\nu = \frac{R}{1 - R}\left(1 + \frac{\gamma_\mu\sigma_\epsilon}{\theta_\mu}\right) M_\nu, \tag{S15}$$

with $\sigma_\epsilon \equiv \epsilon_\mu/\epsilon_m$.

In order to determine the noise levels for a given cost ΔM_ν, we can invert the above equation for R, and then plug into Eq. (S14). In the limit $\phi \ll 1$ we get

$$E(\mathcal{M}, \theta_\mu) \approx \frac{(\gamma_\mu \sigma_\epsilon + \theta_\mu)(\mathcal{M}\theta_\mu^2 + \gamma_\mu^2 \sigma_\epsilon + \gamma_\mu \theta_\mu (1+\mathcal{M}))}{\gamma_\mu \left(\gamma_\mu \sigma_\epsilon + \theta_\mu (1 + \mathcal{M}(1-\theta_\mu))\right)^2}, \tag{S16}$$

where $\mathcal{M} = \Delta M_\nu / M_\nu$.

S.II. THE NOISE-FILTER FORMULATION OF THE MICRORNA REGULATED SYSTEM

A. Realizing miRNA regulation as a noise filtering mechanism

MicroRNA regulation acts directly on mRNA-related fluctuations in protein output production. We can formulate noise regulation as a filtering process by decomposing the protein fluctuations given in Eq. (S10c) as

$$\delta p_R(\omega) = \underbrace{\frac{k_p \delta m_R(\omega)}{d_p - i\omega}}_{s(\omega) - \tilde{s}(\omega)} + \underbrace{\frac{\sqrt{\alpha_p}\eta_p(\omega)}{d_p - i\omega}}_{\delta p_g(\omega)}, \tag{S17}$$

where $\delta p_g(\omega)$ is the ground noise in protein production (the intrinsic noise independent of mRNA fluctuations), and $s(\omega) = \frac{k_p \delta m_R(\omega)}{d_p - i\omega}|_{R=0}$ is the contribution from mRNA noise in the absence of regulation. The perturbation due to regulation, $-\tilde{s}(\omega)$, can then be interpreted in the noise filter framework as an "estimate" of the signal $s(\omega)$, with maximum reduction of the mRNA-induced noise occuring when the estimate is perfect, $\tilde{s}(\omega) = s(\omega)$. Note that $\tilde{s}(\omega)$ and $s(\omega)$ have non-zero correlations, while $\langle(\delta p_g(\omega) s(-\omega))\rangle = \langle(\delta p_g(\omega)\tilde{s}(-\omega))\rangle = 0$. Hence miRNA regulation can compensate for the mRNA noise, but not the ground noise. The normalized error in estimation is defined in the time domain,

$$E = \frac{\langle (s(t) - \tilde{s}(t))^2 \rangle}{\langle s(t)^2 \rangle} \tag{S18}$$

which is equivalent to Eq. (S12) since $\langle (s(t) - \tilde{s}(t))^2 \rangle = (F_R - 1)\bar{p}$ and $\langle s(t)^2 \rangle = (F_0 - 1)\bar{p}$. For a general linear filtering problem we can write the estimated signal as

$$\tilde{s}(\omega) = H(\omega)(s(\omega) + n(\omega)), \tag{S19}$$

where $s(\omega) + n(\omega) \equiv c(\omega)$ is the signal corrupted with noise $n(\omega)$, and $H(\omega)$ is the linear filter function. Eq. (2) of the main text is the inverse Fourier transform of Eq. (S19), showing filtering as a convolution in the time domain,

$$\tilde{s}(t) = \int_{-\infty}^{t} dt' H(t - t')(s(t') + n(t')). \tag{S20}$$

The fact that the upper limit of the integral is t reflects an additional physical constraint on the filter: in the time domain $H(t) = 0$ for all $t < 0$. This enforces causality, namely

that the estimate $\tilde{s}(t)$ can only depend on on the past history ($t' < t$) of the noise-corrupted signal $s(t') + n(t')$. In Fourier space, this translates to the requirement that $H(\omega)$ extended to complex ω can have no poles or zeros in the upper half-plane, $\mathrm{Im}\,\omega > 0$.

For our system, each of terms in Eq. (S19) can be identified using Eqs. (S10b), (S17), and (S19):

$$H(\omega) \equiv \frac{B - A}{A} \tag{S21}$$

$$s(\omega) \equiv \frac{k_p \sqrt{\alpha_m} \eta_m(\omega)}{DA} \tag{S22}$$

$$n(\omega) \equiv -\frac{k_p}{(B - A)D} \left(\sqrt{\alpha_m R/(1 - R)} \eta_b(\omega) - \frac{d_{m\mu}\theta_\mu}{C} \sqrt{\alpha_\mu} \eta_\mu(\omega) \right) \tag{S23}$$

$$c(\omega) \equiv s(\omega) + n(\omega) \tag{S24}$$

with complex functions

$$A \equiv d_m - i\omega, \quad B \equiv c_m - i\omega, \quad C \equiv d_\mu - i\omega, \quad D \equiv d_p - i\omega. \tag{S25}$$

The noise term $n(\omega)$ captures the additional fluctuations due to miRNA-mRNA interactions and it is uncorrelated with the signal $s(\omega)$.

B. Wiener-Kolmogorov optimality of protein output noise level

Wiener-Kolmogorov (WK) theory provides a framework to calculate the optimal causal filter function $H_{\mathrm{wk}}(\omega)$ that minimizes the error E, and thus represents the fundamental limit on how much linearized regulation can reduce mRNA noise. An overview of the WK formalism can be found in Ref. [5]. Here we show how to apply it to the miRNA system.

In order to determine the error function E in Eq. (S18), we need to calculate the variance given in the numerator:

$$\langle (s(t) - \tilde{s}(t))^2 \rangle|_{t=0} = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \left(|H(\omega)|^2 P_{nn}(\omega) + |1 - H(\omega)|^2 P_{ss} \right) \tag{S26}$$

where the power spectra (or cross-spectral density) functions are defined as:

$$2\pi P_{xy}(\omega)\delta(\omega + \omega') = \langle x(\omega) y(\omega') \rangle \tag{S27}$$

for x, $y = s$ or n. The inverse Fourier transform of this expression would give the cross correlations in the time domain, $C_{xy}(t) \equiv \langle x(t') y(t' + t) \rangle$. The WK optimal filter minimizing E for a given $s(t)$, i.e., minimizing Eq. (S26), satisfies

$$H_{\mathrm{wk}}(\omega) = \frac{1}{P_{cc}^{+}(\omega)} \left\{ \frac{P_{cs}(\omega)}{P_{cc}^{-}(\omega)} \right\}_{+}. \tag{S28}$$

The above expression requires two types of decomposition into causal (+) and anti-causal (−) parts, where a causal term in Fourier space has no poles or zeros in the upper complex ω

half-plane, and an anti-causual term is the complex conjugate of a causal function. The first decomposition is multiplicative, where we write $P_{cc} = P_{cc}^{+} P_{cc}^{-}$, a product of causal and anti-causal terms. The second decomposition is additive, denoted with brackets as $\{P_{cs}/P_{cc}^{-}\}_{+}$. Here $\{G(\omega)\}_{+}$ for a function $G(\omega)$ can be calculated by doing a partial fraction expansion of $G(\omega)$ can keeping only terms with no poles in the upper half-plane.

The relevant power spectra are obtained by inserting Eqs. (S22), (S23), (S24) into Eq. (S27):

$$P_{cs} = \frac{{k_p}^2}{|D|^2|A|^2}\alpha_m, \quad P_{cc} = \frac{P_{cs}}{|C|^2 d_m^2}\mathcal{L}\left(\frac{d_m^2|C|^2}{\mathcal{L}} + |A|^2|C|^2 + d_m^2 b_\mu |A|^2\right) \tag{S29}$$

and $P_{ss} = P_{cs}$, $P_{nn} = P_{cc} - P_{cs}$ with dimensionless parameters $b_\mu = \frac{\alpha_\mu(1-R)}{\alpha_m R}(\theta_\mu/\gamma_m)^2$ and $\mathcal{L} = \frac{1-R}{R(1-\theta_\mu)^2}$. To achieve the causal decompositions required for WK equation (S28), P_{cc} can be rewritten as,

$$P_{cc} = \frac{P_{cs}}{|C|^2 d_m^2}\mathcal{L}(\lambda_1 + i\omega)(\lambda_1 - i\omega)(\lambda_2 + i\omega)(\lambda_2 - i\omega) \tag{S30}$$

where $\pm i\lambda_{1,2}$ are the roots of the numerator and given by

$$\lambda_{1,2} = d_m\left(\frac{\Gamma \pm \sqrt{\Gamma^2 - 4\beta}}{2}\right), \ \Gamma = b_\mu + \left(\frac{\gamma_\mu}{\gamma_m}\right)^2 + \frac{1+\mathcal{L}}{\mathcal{L}}, \ \beta = b_\mu + \frac{1+\mathcal{L}}{\mathcal{L}}\left(\frac{\gamma_\mu}{\gamma_m}\right)^2 \tag{S31}$$

Thus, we obtain

$$P_{cc}^{+} = \frac{P_{cs}^{+}}{C d_m}\sqrt{\mathcal{L}}(\lambda_1 - i\omega)(\lambda_2 - i\omega) \tag{S32a}$$

$$\frac{P_{cs}}{P_{cc}^{-}} = \frac{P_{cs}^{+} C^* d_m}{\sqrt{\mathcal{L}}(\lambda_1 + i\omega)(\lambda_2 + i\omega)}. \tag{S32b}$$

where $P_{cs}^{+} = \frac{k_p\sqrt{\alpha_m}}{DA}$. Next, we need to determine the additive decomposition of P_{cs}/P_{cc}^{-}, yielding the casual part

$$\left\{\frac{P_{cs}}{P_{cc}^{-}}\right\}_{+} = \sqrt{\frac{\alpha_m}{\mathcal{L}}}\frac{d_m k_p}{(d_m - d_p)}\left(\frac{d_p + d_\mu}{(d_p + \lambda_1)(d_p + \lambda_2)(d_p - i\omega)} - \frac{d_m + d_\mu}{(d_m + \lambda_1)(d_m + \lambda_2)(d_m - i\omega)}\right). \tag{S33}$$

The optimal WK filter can be calculated from Eq. (S28) by inserting Eqs. (S32a) and (S33), resulting in

$$H_{\mathrm{wk}}(\omega) = \frac{d_m^2(d_m s_m - d_p s_p)}{\mathcal{L} d_m d_p (d_m - d_p)}\frac{(d_\mu - i\omega)(\frac{d_m d_p(s_m - s_p)}{d_m s_m - d_p s_p} - i\omega)}{(\lambda_1 - i\omega)(\lambda_2 - i\omega)} \tag{S34a}$$

$$\text{where} \quad s_m = \frac{d_p(d_m + d_\mu)}{(d_m + \lambda_1)(d_m + \lambda_2)}, \tag{S34b}$$

$$s_p = \frac{d_m(d_p + d_\mu)}{(d_p + \lambda_1)(d_p + \lambda_2)}. \tag{S34c}$$

The filter $H_{\rm wk}(\omega)$ contains two characteristic frequencies λ_1, λ_2. This differs from the functional form of $H(\omega)$ in Eq. (S21) which has a single characteristic frequency d_m^{-1}. To evaluate how close the miRNA-mediated noise regulation can approach WK optimality, we numerically compare the errors $E, E_{\rm wk}$ of both cases respectively. We get the error in estimation in WK optimality using Eq. (S26) in Eq. (S18) with $H(\omega) \to H_{\rm wk}(\omega)$,

$$E_{\rm wk} = \frac{F_{\rm wk} - 1}{F_0 - 1} = \frac{d_m^2}{\mathcal{L}(d_m - d_p)^2}\left((d_m + d_p)\left(\frac{s_m^2}{d_p} + \frac{s_p^2}{d_m}\right) - 4 s_m s_p\right). \tag{S35a}$$

Fig. S1 illustrates the discrepancy between $E_{\rm wk}$ from the above equation and E from Eq. (S12).

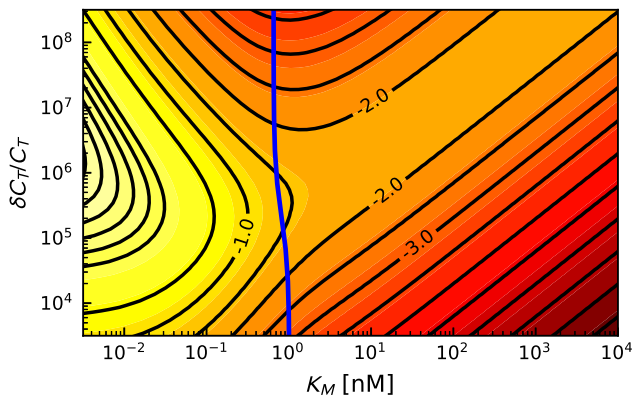

FIG. S1. Contour diagrams of $\log_{10}(E/E_{\rm wk} - 1)$, in terms of Michaelis-Menten constant $K_{\rm M}$ and fractional metabolic cost $\delta C_T/C_T$ for a fixed protein output level $\bar{p}$ in the single target case using the parameters from Table S1. The contour spacing is 0.5, and the blue line follows the most energetically efficient noise reducing $K_{\rm M}^*$ as in Fig. 2 of the main text.

S.III. CALCULATION OF PHASE DIAGRAMS

A. Contour plots, fitness costs, and Michaelis-Menten constant

In the main text, we plot the contour diagram of $\log_{10} E$ from Eq. (S12), describing noise reduction for miRNA-regulated gene expression at a fixed protein output $\bar{p}$. The gene expression parameters are given in Table S1, while the Michaelis-Menten constant $K_{\rm M}$ and miRNA levels $\mu_{\rm tot}$ are taken as free parameters which control both noise reduction and the metabolic costs of regulation. The regulation costs are calculated using Eq. (S15), and the definitions of energetic costs per mRNA and per miRNA respectively,

$$\epsilon_m = \zeta_P L_{m,\rm pre}, \qquad \epsilon_\mu = \zeta_P L_{\mu,\rm pri} \tag{S36}$$

where ζ_P is the energetic need per nucleotide (which we express in units of phosphate bonds hydrolyzed), and $L_{m,\rm pre}$, $L_{\mu,\rm pri}$ are the numbers of nucleotides of precursor mRNA

Parameter	Definition	Value	Reference
d_m	mRNA degradation rate constant	0.11 hr^{-1} per mRNA	[6]
d_p	protein degradation rate constant	0.022 hr^{-1} per protein	[6]
d_μ	microRNA degradation rate constant	0.05 hr^{-1} per miRNA	[7, 8]
$d_{m\mu}$	mRNA degradation through miRISC	0.44 hr^{-1} per complex	
K_M	Michaelis-Menten const. for target degradation	solved for	
$L_{m,\mathrm{pre}}$	precursor mRNA nucleotide number	20,000	[9]
$L_{\mu,\mathrm{pri}}$	primary miRNA nucleotide number	2,500	[10, 11]
ζ_P	P bonds hydrolyzed per nucleotide recycling	2.17 P hr^{-1}	[4]
M_tot	total metabolic rate per cell	3×10^{11} P hr^{-1}	[4]
V	volume of cells	2,000 $\mu\mathrm{m}^3$	[12]
ϕ	d_p/d_m	0.2	
γ_m	$d_m/d_{m\mu}$	0.25	
γ_μ	$d_\mu/d_{m\mu}$	0.114	
$\bar{m}$	mean mRNA copy number	31	[6, 12]
$\bar{p}$	mean protein copy number	200,000	[6, 12]
σ_ϵ	ϵ_μ/ϵ_m	0.125	

TABLE S1. Gene expression parameters (metazoan) used in calculating the phase diagrams. Typical (or estimated) values from the corresponding references are taken such as median values over a distribution of genes. When an extensive distribution is not present, a reasonable estimation is made compatible with the references. We have not found a value for $d_{m\mu}$ in the literature, however, target mRNAs degrade faster in the complex, and hence we assumed the ratio to be within an order of magnitude. In order to convert the copy number units to molar units of mRNA numbers and K_m, we divide these values by the cell volume V and Avagadro constant.

and primary miRNA. In principle, the costs associated with the miRISC can be larger than ϵ_μ. Thus, our estimations can be seen as a lower bound. Finally, we relate metabolic costs to a fitness disadvantage, $s_c \sim \Delta M_\nu / M_\mathrm{tot}$ where M_tot is the (mean) total metabolic rate of a cell, as discussed in the main text.

B. Minimal noise for the same metabolic cost

The contour plots indicate a minimum E for a given metabolic cost $\sim \Delta M_\nu$ at a precise K_M value. This value remains in a narrow range as a function of metabolic costs as shown in the Fig. 2 of the main text. Taking the limit $\phi \ll 1$, we can estimate this value by minimizing Eq. (S16) with respect to $\theta_\mu(K_\mathrm{M})$. Among multiple solutions, there is a single physical one, i.e., $0 \leq \theta_\mu(K_\mathrm{M}^*) \leq 1$, from which we can infer K_M^* using Eq. (S7). For high $\mathcal{M}$ and $\gamma_\mu \ll 1$, this yields the Eq. 4 of the main text,

$$K_\mathrm{M}^*(\Delta M_\nu, \bar{p}) \approx \bar{m} \gamma_\mu^{-1} \sigma_\epsilon^{-1/2}. \tag{S37}$$

This approximation holds well for the parameters in Table S1, with only 0.6% discrepancy at high $\mathcal{M}$. For $\sigma_\epsilon \ll 1$ we observe that $E(K_{\rm M}^*) \sim \mathcal{M}^{-1}$, which relates noise reduction to metabolic and hence to the fitness costs, i.e., $s_c \propto -M_\nu/E(K_{\rm M}^*)$.

S.IV. ESTIMATING DISSOCIATION CONSTANTS $K_{\rm D}$ FOR SEED SEQUENCES OF VARYING LENGTH

To estimate a range of possible $K_{\rm D}$ values for miRNAs with different seed sequences, we started with a collection of 9990 human miRNA seed sequences of length 7 nucleotides (nt) taken from the TargetScanHuman database, ver. 7.1 [14]. For each sequence, we used the RNAcofold algorithm from the ViennaRNA package (ver. 2.3) to calculate the free energy ΔG^0 at $T = 298°$K for binding to a complementary target sequence [13]. To mimic the effects of unpaired nucleotides flanking the seed-target complex, we added two random unpaired nucleotides at both the beginning and end of the seed/target to make the total length 11 nt. The algorithm was run with the default dangling end energy option. Once ΔG^0 is known, $K_{\rm D}$ can be found using the relation $\Delta G^0 = k_B T \ln(K_{\rm D}/[1M])$.

To validate the predictions of the algorithm, we checked it on the seed sequence GAGGUAG, for which experimental $K_{\rm D}$ values are available for fruit fly and mouse siRNA-target complexes [15]. Depending on the dangling ends, the algorithm predicted a range of $K_{\rm D}$ with mean 63 pM (95% confidence interval: 7 to 238 pM). This compares well with the experimentally measured ranges for different seed-matched targets: 4 to 210 pM for fly, 13 to 26 pM for mouse [15].

A histogram of the results for all the seed sequences in the dataset is shown in Fig. S2A. The distribution has 8 peaks, roughly corresponding to the fact that the seed-target complex can have between 0 and 7 CG pairs. The more CG pairs relative to AU pairs, the stronger the binding and the lower the $K_{\rm D}$. To model the effect of having a 6 nt seed, in Fig. S2B the calculation was redone after randomly deleting 1 nucleotide from each original seed. Similarly Fig. S2C shows the results for 5 nt seeds, after randomly deleting 2 nucleotides per seed. As expected, shorter seeds lead to overall weaker binding, shifting the distribution to larger $K_{\rm D}$ values.

S.V. ANALYSIS OF EXPERIMENTAL MICRORNA SYSTEMS

In this section we use our theoretical framework to analyze data from two earlier experimental works on microRNA-mediated noise regulation. As mentioned in Sec. S.I., the basic mathematical framework for miRNA-regulated protein production, which is the starting point for our theory, has already been validated in Ref. [2] in a variety of experimental scenarios. Here we show another aspect of the analysis, interpreting the experimental data via the optimal noise filter formulation of the theory. We focus on two systems, both involving endogenous miRNAs in living cells interacting with binding sites on 3'UTR regions from endogenous genes, with the latter coupled to fluorescent reporters that allow direct measurements of regulation strength and noise reduction. The genes are: i) *Lats2* in mouse

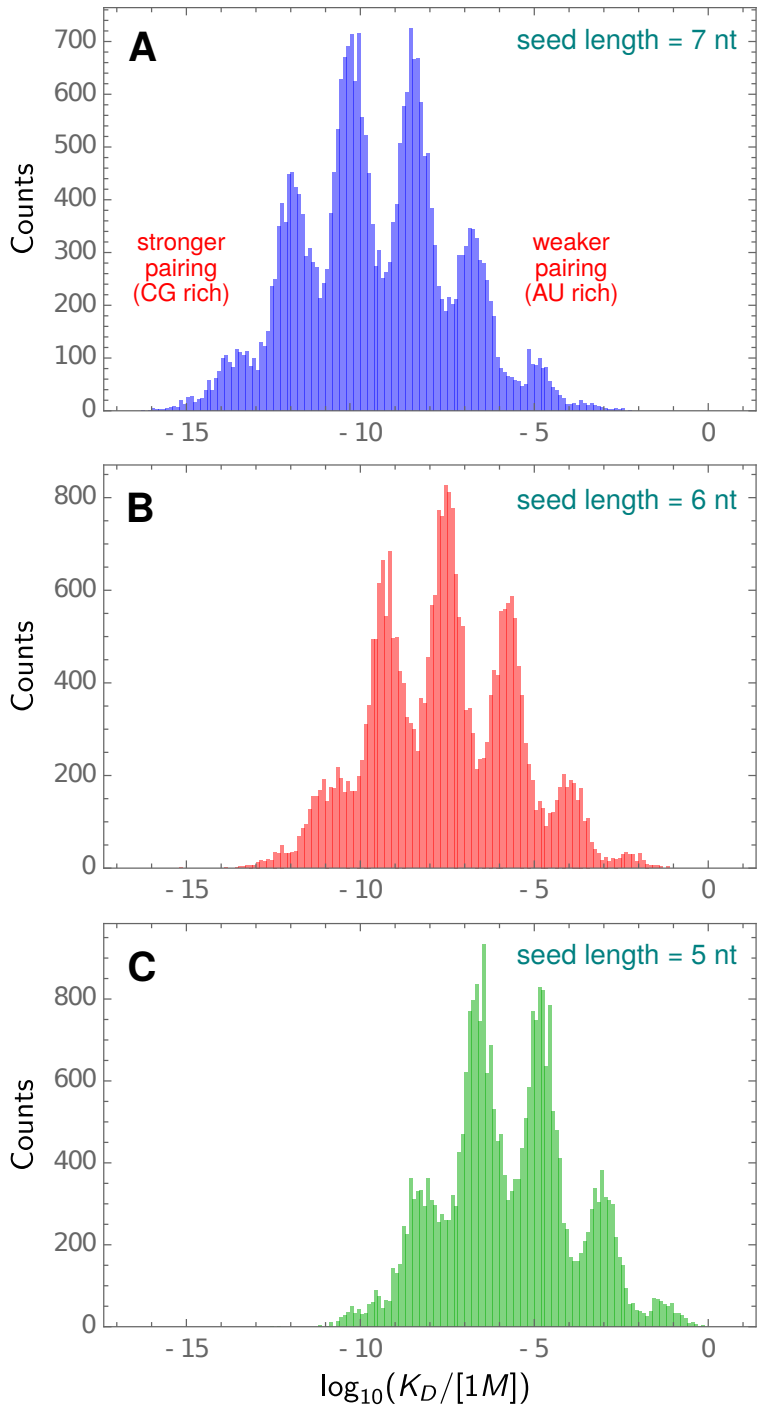

FIG. S2. Distribution of K_{D} values estimated for a different miRNA seed sequences with length: A) 7 nt; B) 6 nt; C) 5 nt. The calculation used the RNAcofold algorithm of the ViennaRNA package [13].

embryonic stem cells (mESCs) [2]; ii) *sens* in *Drosophila* wing disc cells [16].

Let us start with an overview of the analysis procedure, before turning to specific details for the two systems. In each experiment there is a comparison between a miRNA-regulated system, and one without regulation, where the 3'UTR target sites are modified to prevent miRNA binding. This allows us to extract the error E from the relative noise levels in the regulated versus unregulated case, as defined in Eq. (S12), as well as the regulation strength R. The experimental measurements are carried out over a population of cells where the protein expression levels (as quantified by the fluorescent reporter) vary, and we will use X to denote the measure of protein levels in a cell, with X corresponding to fluorescence intensity (in arbitrary units) in the *Lats2* case, and to protein numbers in the *sens* case (where the calibration between intensity and copy numbers was carried out). The noise filter characteristics in a given cell will depend on $X \propto \bar{p} \propto \bar{m}$, and hence the experimental values of $E^{\text{exp}}(X)$ and $R^{\text{exp}}(X)$ will both vary with X. In our analysis, we fit $E^{\text{exp}}(X)$ to Eq. (S14), our analytical result in the limit where the protein degradation rate constant is much smaller than the mRNA degradation rate coefficient ($\phi \ll 1$). Plugging Eq. (S7) and $R^{\text{exp}}(X)$ into Eq. (S14), we get the following theoretical expression relating the two measured quantities:

$$E^{\text{exp}}(X) = \frac{(1 - R^{\text{exp}}(X))\left(1 + \frac{R^{\text{exp}}(X)}{\gamma_\mu}\left(\frac{1}{\widehat{K}_{\text{M}}/X+1}\right)\right)}{\left(1 - R^{\text{exp}}(X)\left(\frac{1}{\widehat{K}_{\text{M}}/X+1}\right)\right)^2}. \tag{S38}$$

where we define a rescaled Michaelis-Menten constant $\widehat{K}_{\text{M}} = K_{\text{M}}X/\bar{m}$, with $X/\bar{m}$ being the constant scaling factor between X and mRNA number $\bar{m}$. Eq. (S38) has only two unknowns, γ_μ and $\widehat{K}_{\text{M}}$, which can be estimated via fitting to the experimental data for $E^{\text{exp}}(X)$ and $R^{\text{exp}}(X)$ as a function of X. Once $\widehat{K}_{\text{M}}$ is determined from the fit, we can use main text Eq. [4] for K^*_{M} to find the ratio of actual to optimal Michaelis-Menten constants, $K_{\text{M}}/K^*_{\text{M}} = \widehat{K}_{\text{M}}\gamma_\mu\sigma_\epsilon^{1/2}/X$, which is plotted in main text Fig. 4. Note that this ratio scales inversely with expression level, $K_{\text{M}}/K^*_{\text{M}} \propto 1/X$, and thus cells with different X will be closer or further away from optimality. As discussed in the main text, the results demonstrate that across a physiological range of protein expression levels, both experimental systems exhibit K_{M} roughly within an order of magnitude of K^*_{M}.

miRNA regulation of *Lats2* in mouse embryonic stem cells [2]: The regulation strength R is obtained by the relative expression levels in regulated and unregulated cases, i.e., $R = 1-\bar{p}_R/\bar{p}_0$. This was determined via a dual reporter system and is almost constant at $R^{\text{exp}}(X) \approx 0.8$ for all expression levels X (Fig. S3A), quantified in terms of the fluorescent marker (mCherry) intensity X. The protein expression noise in these experiments was reported in terms of the coefficient of variation,

$$\eta_\lambda = \sqrt{\frac{\langle \delta p^2 \rangle_\lambda}{\bar{p}^2}} \tag{S39}$$

where the subscript $\lambda = 0, R$ applies to unregulated and regulated cases respectively. These

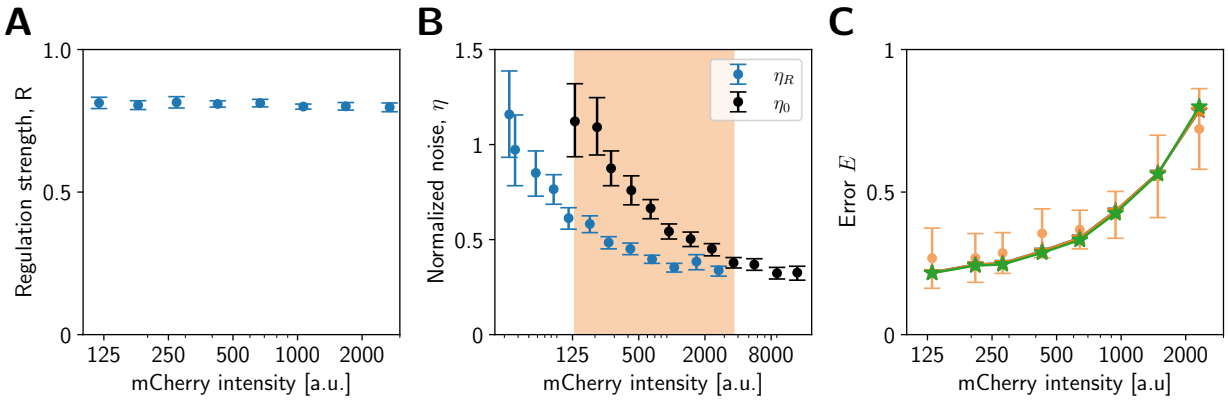

FIG. S3. Analysis of the experimental mCherry expression from the *Lats2*-microRNA system [2]. (A) Experimentally determined regulation strengths R with varying mCherry intensity. (B) Experimental data for noise levels in unregulated (η_0, black) and regulated (η_R, blue) mCherry expression. (C) We span mCherry intensity levels, allowing a direct comparison (within the orange shaded region of (B)) of the noise levels, and transform them to error E (orange) using Eq.(S41) and numerical interpolation. The solid lines (multi-color) with stars as points display best-fit curves for $\gamma_\mu = 0.001, 0.01, 0.114$. These three curves closely overlap, so are indistinguishable.

are shown in Fig. S3B as blue (regulated data), black (unregulated data). Our theory can be related to the experimental data by estimating the error value,

$$E = \frac{F_R - 1}{F_0 - 1} \approx \frac{F_R}{F_0} \tag{S40}$$

where $F_\lambda = \langle \delta p^2 \rangle_\lambda / \bar{p}$ for $\lambda = 0, R$ (regulated and unregulated) and the approximation holds when $F_R, F_0 \gg 1$, which is generally true for protein expression [17]. Thus we can estimate the experimental error $E^{\text{exp}}(X)$ from the reported data via

$$E^{\text{exp}}(X) = \frac{\eta_R^2(X)}{\eta_0^2(X)}, \tag{S41}$$

through which we obtain $E^{\text{exp}}(X)$ as a function of X in the overlapping region (shaded region in Fig. S3B). Since $\eta_R(X)$ and $\eta_0(X)$ are not measured at exactly the same X values, we use interpolation to match the values, and show the results for the error $E^{\text{exp}}(X)$ in Fig. S3C (orange points). To calculate the error bars, we have used the appropriate mean and standard deviation error propagation rules from the original data.

To apply Eq. (S38), we use the average $R^{\text{exp}}(X)$ value and perform a chi-squared fitting to $E^{\text{exp}}(X)$ using its mean and standard deviations. The quality of fitting is relatively insensitive to $\gamma_\mu = d_\mu / d_{m\mu}$. Since it is assumed γ_μ should be smaller than 1 (degradation of the miRNA-mRNA complex is faster than miRNA alone), but we do not have good literature estimates for $d_{m\mu}$, we take a biologically plausible range of possible γ_μ values $\gamma_\mu = 0.001 - 0.114$. The results are shown in Fig. S3C, and curves for different γ_μ values are essentially indistinguishable, all providing good fits to the experimental data.

In Fig. 4A of the main text, we plot the inferred Michaelis-Menten constants normalized by the optimal Michaelis-Menten value, $K_{\rm M}/K_{\rm M}^* = \widehat{K}_{\rm M}(\gamma_\mu)\gamma_\mu\sigma_\epsilon^{1/2}/X$, where σ_ϵ is the ratio of metabolic costs of microRNA and mRNA transcripts. The points represent fits using the Table S1 values of γ_μ and σ_ϵ, and the colored regions show the uncertainty due to possible ranges of γ_μ and σ_ϵ. The ratio $K_{\rm M}/K_{\rm M}^*$ depends very weakly on the choice of γ_μ, so the range $\gamma_\mu = 0.001 - 0.114$ does not significantly affect the estimate (in fact even extending the range to $\gamma_\mu < 0.001$ leaves the uncertainty region for $K_{\rm M}/K_{\rm M}^*$ essentially unchanged). For σ_ϵ we chose the range $\sigma_\epsilon = 0.0014 - 0.5$ as follows. The energetic costs of miRNA and mRNA transcription were introduced in Sec. S.III, with $\sigma_\epsilon = L_{\mu,\rm pri}/L_{\rm m,pre}$. The average estimated lengths of pri-miRNA $L_{\mu,\rm pri} = 2,500$ and pre-mRNAs $L_{\rm m,pre} = 20,000$ lead to $\sigma_\epsilon = 0.125$ (Table S1). We can narrow our search knowing that $L_{\rm m,pre} \approx 50,000$ for the Lats2 gene [18]. However, the pri-miRNA transcripts are largely unknown due to fast transcription from primary to precursor structure. Moreover, a single pri-miRNA can lead to multiple copies of pre-miRNA. Thus, despite proposed methods [19, 20], we lack this information for mouse microRNAs. Nevertheless, considering the lengths of pre-miRNAs, which are about 70 nucleotides [21], gives an approximate lower bound: $L_{\mu,\rm pri} \geq L_{\mu,\rm pre} = 70$, which yields to $\sigma_\epsilon = 0.0014$ (note that this is a strict underestimation). Since the energetic costs of assembling miRNA are expected to be smaller than mRNA, we use $\sigma_\epsilon = 0.5$ as an upper bound. The lowest and highest $K_{\rm M}/K_{\rm M}^*$ values (boundaries of colored region in Fig. 4A) correspond to the results for lowest and highest values of both γ_μ and σ_ϵ.

Finally, as a consistency check, we can attempt to approximately convert the value of the inferred $K_{\rm M}$ into units of molar concentration. For this, we need a conversion factor between mCherry intensities and mRNA concentration. In Fig. 4B of Ref. [2], the probability density of mRNA levels in the mESC transcriptome is correlated with the range of mCherry fluorescence intensities. The peak value (most likely mRNA level in RPKM units) corresponds to intensity $X^\dagger = 510$. While we do not know the peak value for mESC cells in terms of copy numbers, we can use data from mouse fibroblasts [6] as a rough proxy. There the peak occurs around a copy number $\bar{m}^\dagger = 8$ (the most likely number of mRNAs from a single gene). Assuming a cell volume of 2,000 μm^3, this amounts to molar density $[m]^\dagger_{\rm fibroblast} = 0.0066$ nM. If we take mESC concentration at the peak $[m]^\dagger_{\rm mESC}$ to be similar, then the conversion factor is $[m]^\dagger_{\rm mESC}/X^\dagger = 1.29 \times 10^{-5}$nM/mCherry intensity. We can obtain $K_{\rm M}$ in terms of molar concentration using $K_{\rm M} = \widehat{K}_M[m]/X$ and the conversion factor, which yields $K_{\rm M} = 8.35, 0.86, 0.1$ nM respectively for the possible parameter values $\gamma_\mu = 0.001, 0.01, 0.114$. For comparison, $K_{\rm M}$ for a miRNA system in mouse fibroblasts was measured to be 0.1 nM [15], so our estimated $K_{\rm M}$ numbers seem plausible (particularly at $\gamma_\mu = 0.114$, the value from Table S1 which we used for most of our calculations).

miRNA regulation of *sens* in *Drosophila* wing disc cells [16]: These experiments use a similar technique to the one described above, with reporter proteins attached to the *sens* 3'UTR used to investigate the noise regulation mediated by miRNAs. Comparison of noise levels is made between proteins having 3'UTR sequences with and without miR-9a binding sites (mCherry Sens wild type and mCherry Sens mutant). We obtained the regulation strengths $R^{\rm exp}(X)$ by using the raw data for relative expression levels of mCherry Sens wild

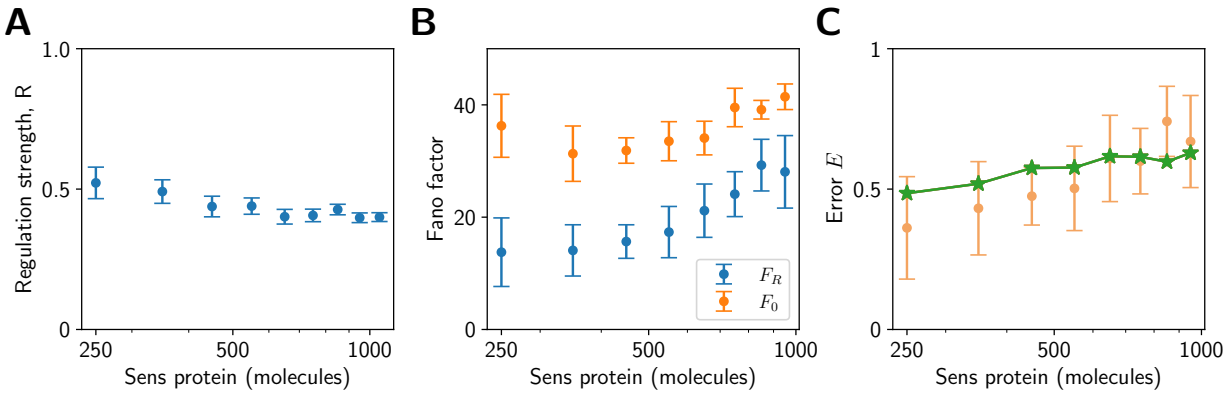

FIG. S4. Analysis of the experimental *sens*-miRNA system in *Drosophila* [16]. (A) Experimentally determined regulation strengths R with varying Sens protein numbers. (B) Experimental data for the Fano factors of the unregulated (F_0, orange) and regulated (F_R, blue) cases. (C) Experimental results for the error E (orange), compared to best-fit theoretical curves (multi-color lines with stars correspond to $\gamma_\mu = 0.001$, 0.01, 0.114, and appear indistinguishable).

type and sfGFP Sens mutant proteins. This is plotted in Fig. S4A. The protein expression noise levels are reported in terms of Fano factors F_0 and F_R, and we binned the data along different expression levels (Fig. S4B). E^{exp} can then be directly obtained from:

$$E^{\text{exp}} = \frac{F_R - 1}{F_0 - 1}, \tag{S42}$$

which is shown in Fig. S4C. We then apply a fitting procedure identical to the one described in the *Lats2* example.

Main text Fig. 4B plots the inferred ratio $K_{\text{M}}/K_{\text{M}}^* = \widehat{K}_{\text{M}}(\gamma_\mu)\gamma_\mu\sigma_\epsilon^{1/2}/X$, with points representing fits using the parameter values from Table S1. The colored uncertainty region corresponds to varying $\gamma_\mu = 0.001 - 0.114$ and $\sigma_\epsilon = 0.014 - 0.5$. In this case, the lower bound for σ_ϵ is obtained using the pre-mRNA length of the *sens* gene $L_{\text{m,pre}} \approx 5,000$ [22] and $L_{\mu,\text{pri}} \geq L_{\mu,\text{pre}} = 70$, yielding $\sigma_\epsilon = 0.014$.

[1] D. T. Gillespie, J. Chem. Phys. **113**, 297 (2000).

[2] J. M. Schmiedel, S. L. Klemm, Y. Zheng, A. Sahay, N. Blüthgen, D. S. Marks, and A. van Oudenaarden, Science **348**, 128 (2015).

[3] S. Mukherji, M. S. Ebert, G. X. Zheng, J. S. Tsang, P. A. Sharp, and A. Van Oudenaarden, Nat. Genet. **43**, 854 (2011).

[4] M. Lynch and G. K. Marinov, Proc. Natl. Acad. Sci. **112**, 15690 (2015).

[5] D. Hathcock, J. Sheehy, C. Weisenberger, E. Ilker, and M. Hinczewski, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 16 (2016).

[6] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, Nature **473**, 337 (2011).

[7] M. J. Marzi, F. Ghini, B. Cerruti, S. De Pretis, P. Bonetti, C. Giacomelli, M. M. Gorski, T. Kress, M. Pelizzola, H. Muller, *et al.*, Genome research **26**, 554 (2016).

[8] E. R. Kingston and D. P. Bartel, Genome Res. **29**, 1777 (2019).

[9] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland Science, 2015).

[10] G. Song and L. Wang, PloS One **3**, e3574 (2008).

[11] H. K. Saini, A. J. Enright, and S. Griffiths-Jones, BMC Genom. **9**, 1 (2008).

[12] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, Nucleic Acids Res. **38**, D750 (2010), http://bionumbers.hms.harvard.edu/.

[13] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, Algorithms Mol. Biol. **6**, 1 (2011).

[14] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, eLife **4**, e05005 (2015).

[15] L. M. Wee, C. F. Flores-Jasso, W. E. Salomon, and P. D. Zamore, Cell **151**, 1055 (2012).

[16] R. Giri, D. K. Papadopoulos, D. M. Posadas, H. K. Potluri, P. Tomancak, M. Mani, and R. W. Carthew, eLife **9**, e53638 (2020).

[17] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. Van Oudenaarden, Nature genetics **31**, 69 (2002).

[18] J. A. Blake, R. Baldarelli, J. A. Kadin, J. E. Richardson, C. L. Smith, and C. J. Bult, Nucl. Acids Res. **49**, D981 (2021).

[19] J. Qian, Z. Zhang, J. Liang, Q. Ge, X. Duan, F. Ma, and F. Li, Genomics **97**, 294 (2011).

[20] K. Bedi, M. T. Paulsen, T. E. Wilson, and M. Ljungman, NAR genomics and bioinformatics **2**, lqz014 (2020).

[21] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim, The EMBO journal **21**, 4663 (2002).

[22] L. S. Gramates, J. Agapite, H. Attrill, B. R. Calvi, M. A. Crosby, G. Dos Santos, J. L. Goodman, D. Goutte-Gattat, V. K. Jenkins, T. Kaufman, *et al.*, Genetics **220**, iyac035 (2022).