Monoallelic expression as an emergent property of stochastic gene regulation

Jorge Aurelio Menéndez

CoMPLEX Mini-Project #1
Friday, 22nd January 2016

Contents

1. Introduction

Random *monoallelic expression* (MAE) is the phenomenon whereby, in diploid organisms, one allele of a gene is significantly more expressed than the other, usually by a factor of 10-50. [5, 6, 9, 17] Importantly, as opposed to genomic imprinting, the selection of which allele is expressed is random. [5, 2] Some cells will express one allele while others express another, and in many cases others will simply express both (*biallelic expression*). [9, 10, 6, 12] Genome-wide searches of monoallelic expression have found that 5-10% of genes in the human genome are susceptible to monoallelic expression [9], and it has been argued that this is likely a lower bound. [17]

Such an expression pattern can be biologically advantageous in a number of domains. For example, in X-inactivation, most X-linked genes are transcribed from only one of the two X-chromosomes in females. This allows for dosage compensation of the proteins coded for by these genes, as the presence of two X-chromosomes in females would otherwise lead to doubling of their expression relative to males (who have only one X-chromosome). [5] Looking at autosomal genes, two canonical examples of monoallelic expression are olfactory receptor gene [3] and B-cell receptor gene [16] expression. Because autosomal monoallelic expression is not chromosome-coordinated (i.e. the allelic choice of a given MAE gene is independent of the allelic choices of other MAE genes on the same chromosome) [9], monoallelic expression can massively increase combinatorial diversity [17, 14, 1], something certainly desirable and even necessary for the olfactory system (which needs to be able to distinguish many different odors) and the immune system (which needs to be able to respond to many different antigens). In most other cases, however, autosomal random monoallelic expression doesn't have a clear purpose. Indeed, it is often the case that a gene monoallelically expressed in one cell is biallelically expressed in another cell, suggesting that monoallelic expression is simply a feature of certain cell lines rather than a requirement for survival or proper functioning. [10, 12]

Such cases are the focus of the present study. Specifically, we asked whether a simple stochastic model of gene expression could account for the patterns and prevalence of monoallelic expression found in the literature. Importantly, the model only incorporates regulatory connections between genes in a genetic network. This is quite a novel approach in that few previous studies have explored the possibility that monoallelic expression could arise solely from gene regulation (although see the "genetic control hypothesis" of Savova *et al.*, 2013). We hypothesized that certain properties of a gene's regulatory connections would make it more susceptible to a monoallelic mode of expression, i.e. that all monoallelically expressed genes would show a similar regulation profile. A mathematical model of stochastic gene expression and regulation allowed us to test this hypothesis on theoretical grounds. Subsequent efforts should seek to test it empirically, although it is quite difficult to measure and determine the properties and mechanisms of gene expression networks at the molecular scale.

I begin by outlining the mathematical model used and our simulation methods. I then go on to present our findings: first, confirming that our model is appropriate, second, showing that indeed stochastic gene regulation by itself can lead to monoallelic expression, and, third, highlighting some of the features of the monoallelism demonstrated by our model. I finish by discussing the implications of these findings and how they fit in with the literature, suggesting future theoretical and empirical research directions.

2. Methods

2.1. **Gene expression model.**

We model diploid gene expression in a gene network as a stochastic birth-death process [7, 8], where the state of the system is defined by the number of mRNA transcripts and protein molecules produced from each allele of each gene in the network. The birth processes are thus transcription and translation, respectively, and the death processes mRNA and protein degradation. This gives us the following reaction scheme:

$$\begin{array}{lll}
\text{(a)} & r_i^{(a)} \xrightarrow{k_{Ri}} r_i^{(a)} + 1 & \text{transcription} \\
\text{(b)} & p_i^{(a)} \xrightarrow{k_P r_i^{(a)}} p_i^{(a)} + 1 & \text{translation} \\
\text{(c)} & r_i^{(a)} \xrightarrow{\gamma_R r_i^{(a)}} r_i^{(a)} - 1 & \text{mRNA degradation} \\
\text{(d)} & p_i^{(a)} \xrightarrow{\gamma_P p_i^{(a)}} p_i^{(a)} - 1 & \text{protein degradation}
\end{array}$$

where $r_i^{(a)}$ denotes the number of mRNA transcripts transcribed from the ath allele of the ith gene, and $p_i^{(a)}$ denotes the number of protein molecules translated from these transcripts. Note that the transcription rate k_R is subscripted with an i, indicating that it is specific to each gene. As we will see below, this is necessary to allow for regulatory connections between genes.

Interpreting the above reaction rates as probabilities, this system yields the following two master equations [7]:

$$\begin{aligned}\frac{d\,\mathrm{Pr}(r_i^{(a)})}{dt} &= k_{Ri}\,\mathrm{Pr}(r_i^{(a)}-1)+\gamma_R(r_i^{(a)}+1)\,\mathrm{Pr}(r_i^{(a)}+1)\\ &\quad-(k_{Ri}+\gamma_R r_i^{(a)})\,\mathrm{Pr}(r_i^{(a)})\end{aligned}\tag{1}$$

$$\begin{aligned}\frac{d\,\mathrm{Pr}(p_i^{(a)})}{dt} &= k_P r_i^{(a)}\,\mathrm{Pr}(p_i^{(a)}-1)+\gamma_P(p_i^{(a)}+1)\,\mathrm{Pr}(p_i^{(a)}+1)\\ &\quad-(k_P r_i^{(a)}+\gamma_P p_i^{(a)})\,\mathrm{Pr}(p_i^{(a)})\end{aligned}\tag{2}$$

which, put together, lead to the full master equation for an arbitrary gene network with N genes:

$$\begin{aligned}\frac{d\,\mathrm{Pr}(\langle\vec{r},\vec{p}\rangle)}{dt} =&\sum_{i=1}^{N} k_{Ri}\,\mathrm{Pr}(r_i^{(a)}-1)+\gamma_R(r_i^{(a)}+1)\,\mathrm{Pr}(r_i^{(a)}+1)\\ &\quad-(k_{Ri}+\gamma_R r_i^{(a)})\,\mathrm{Pr}(r_i^{(a)})\ +\\ &k_P r_i^{(a)}\,\mathrm{Pr}(p_i^{(a)}-1)+\gamma_P(p_i^{(a)}+1)\,\mathrm{Pr}(p_i^{(a)}+1)\\ &\quad-(k_P r_i^{(a)}+\gamma_P p_i^{(a)})\,\mathrm{Pr}(p_i^{(a)})\end{aligned}\tag{3}$$

The ordered pair $\langle\vec{r},\vec{p}\rangle$ denotes the current state of the system, where $\vec{r}=(r_1^{(1)},r_1^{(2)},...,r_N^{(1)},r_N^{(2)})$ and $\vec{p}=(p_1^{(1)},p_1^{(2)},...,p_N^{(1)},p_N^{(2)})$.

Gene expression regulation is incorporated into the system through the gene-specific transcription rate k_{Ri} and a regulation matrix R. For a network with N genes, R is an $N\times N$ matrix, where R_{ij} indicates whether the protein expressed by the jth gene regulates transcription of the ith gene: $R_{ij}=0$ for no regulation, 1 for positive regulation, and -1 for negative regulation. k_{Ri} depends on $\vec{p}$ and R through the following equation:

$$k_{Ri}=k_{leak}+k_R^{max}\prod_{j=1}^{N}\phi_i(p_j^{(1)}+p_j^{(2)},K_{ij}^{(a)},n)\tag{4}$$

where ϕ_i is the Hill function with the conditions defined by the ith row of R:

$$\phi_i(p_j^{total}, K_{ij}^{(a)}, n) = \begin{cases} \frac{1}{1+\left(\frac{p_j^{total}}{K_{ij}^{(a)}}\right)^n} & \text{if } R_{ij} < 0 \\ 1 - \frac{1}{1+\left(\frac{p_j^{total}}{K_{ij}^{(a)}}\right)^n} & \text{if } R_{ij} > 0 \\ 1 & \text{if } R_{ij} = 0 \end{cases}$$

Here, n is the Hill coefficient and $K_{ij}^{(a)}$ is the dissociation constant ($\propto$inverse of binding affinity) for the interaction between the jth regulator protein and the locus of the ath allele of the ith gene. Since we only considered homozygotes in the present study, $K_{ij}^{(1)} = K_{ij}^{(2)}$, reflecting the identical nucleotide sequences at the loci of each allele. [18]

Note that the amount of protein regulating each allele is the same for both alleles of the same gene (hence the p_j^{total} in the definition of ϕ_i, where $p_j^{total} = p_j^{(1)} + p_j^{(2)}$). This falls out of the simple fact that if protein j regulates gene i, it regulates the expression of both alleles of that gene, irrespective of its originating allele. Thus, the amount of protein regulating either allele of gene i should be summed over both of its originating alleles. [18]

2.2. **Model simulation.** Finding analytical solutions to master equations can be extremely difficult and often impossible. However, rigorous approximations exist. In line with other studies involving modelling gene expression, I simulated the above model using the Gillespie algorithm. This algorithm is formally derived from the assumptions underlying the master equation approach to modelling birth-death processes, such that it provides an exact simulation of the dynamics of the system. [8]

I now go on to outline the algorithm in pseudo-code, following the notation used in section 2.1. The algorithm was implemented in MATLAB.

(1) *Initialization.* Create variables $r_1^{(1)}, r_1^{(2)}, ..., r_N^{(1)}, r_N^{(2)}, p_1^{(1)}, p_1^{(2)}, ..., p_N^{(1)}, p_N^{(2)}$, and **sim_time**. Initialize all to 0.

(2) *Compute probabilities.* Compute the probability of each of the possible reactions: transcription, translation, mRNA degradation, and protein degradation, for each allele of each gene. For a network consisting of N genes, this yields a vector of (4 reactions)×(2 alleles)×N genes$= 8N$ probabilities. The probabilities are equal to the reaction rates of each reaction (see the reaction scheme above). For the ath allele of the ith gene, these are:
   (a) Pr(transcription) $= k_{Ri}$
   (b) Pr(translation) $= k_P r_i^{(a)}$
   (c) Pr(mRNA degradation) $= \gamma_R r_i^{(a)}$
   (d) Pr(protein degradation) $= \gamma_P p_i^{(a)}$

where k_{Ri} is computed using equation (4). Let $\vec{\rho} = (\rho_1, ..., \rho_M)$ be the resulting vector of probabilities, where $M = 8N$.

(3) *Sample time until next reaction.* Randomly sample the amount of time until the next reaction occurs. Making the assumption that reaction frequency is Poisson distributed, the time between reactions is sampled from an exponential distribution with mean $\sum \rho_k$. Add the sampled amount of time to variable **sim_time**.

(4) *Sample reaction.* Sample the reaction to occur at the current time step. This is done by sampling a random number between 0 and 1 from the uniform distribution and seeing where it falls within a partitioned space in which each partition corresponds to a different reaction. The respective sizes of these partitions are scaled to the relative probabilities of their corresponding reactions (saved in $\vec{\rho}$). If the random number falls in the partition corresponding to reaction X, then reaction X is selected to occur. Because the size of any given partition is scaled to the relative probability of its corresponding reaction, the probability of a uniformly distributed number falling in that partition is equal to the relative probability of that reaction. Thus, reactions sampled in this way are sampled according to their computed probability distribution.

(5) *Update system state.* Depending on the selected reaction X, update the appropriate variables:

   (a) If $X =$ *transcription* of allele a of gene i, $r_i^{(a)} = r_i^{(a)} + 1$

   (b) If $X =$ *translation* of mRNA from allele a of gene i, $p_i^{(a)} = p_i^{(a)} + 1$

   (c) If $X =$ *mRNA degradation* of the mRNA from allele a of gene i, $r_i^{(a)} = r_i^{(a)} - 1$

   (d) If $X =$ *protein degradation* of the protein coded for by allele a of gene i, $p_i^{(a)} = p_i^{(a)} - 1$

(6) *Save protein concentrations to timecourse.* Save the current amounts of each protein ($\vec{p}$) and the current time (**sim_time**).

(7) *Repeat.*Continue until variable **sim_time** reaches 20×protein half-life, where protein half-life $= \frac{\log 2}{\gamma_P}$.

2.3. **Simulation parameters.** In all the simulations reported below (with the exception of the simulation time parameter in experiment 5), the following parameter values were used. Apart from the Hill coefficient n and the simulation time, these were taken from Thattai & van Oudenaarden (2001).

Protein half-life is computed from γ_P by $\frac{\log 2}{\gamma_P} = 3600$ secs. $= 1$ hr. It is worth noting that the Hill coefficient used was $n = 1$. This greatly reduces non-linear effects in regulation, making our rich pattern of results particularly striking.

Table 1. Simulation parameters

Parameter	**Value**
k_R^{max}	.01 s^{-1}
k_{leak}	.0001 s^{-1}
k_P	.1 s^{-1}
γ_R	5.80×10^{-3} s^{-1}
γ_P	1.93×10^{-4} s^{-1}
n	1
simulation time	20×protein half-life

3. Results

3.1. **Replication of previous work.** To ensure that our algorithm for simulating gene expression was working correctly, we first tried replicating the behavior of the Thattai & van Oudenaarden (2001) haploid gene expression model that ours is based on. We modified our diploid gene expression model to only include one allele of each gene, everything else kept exactly the same.

One of their primary findings was that autoregulation leads to a decrease in gene expression noise. To replicate this finding, we constructed eight single-gene networks, where the protein expressed by the gene negatively regulated its transcription (i.e. $R = -1$, a 1×1 matrix). Critically, we manipulated the strength of this repression by changing the dissociation constant of the gene in each network. As is clearly demonstrated by figure 1 below, the stronger the autoregulation (the lower the dissociation constant), the narrower the range of expression levels across simulation runs. Indeed, Fano noise decreased with decreasing dissociation constant (table 2).

A second finding of this study was that a two-gene network composed of two mutually repressing proteins has the dynamics of a bistable switch: two steady states exist, and the system can switch from one to the other. This leads to the shallow valley between the two peaks observed in figure 2. Figure 3 replicates the dynamics of the three-gene network simulated by Thattai & van Oudenaarden (2001), consisting of a feed-forward cascade of three genes where each one represses the next.[1]

3.2. **Experiments 1 & 2: monoallelic expression in random gene networks.** To examine the relationship between gene regulation and monoallelic expression, we began by generating random networks and simulating them. Each network was assigned a random number of genes between 1 and 10, and each element of the regulation matrix R was assigned 0, 1, or -1, with equal probability of

[1] It should be noted that the dissociation constants of the networks simulated by Thattai & van Oudenaarden (2001) were not provided. Thus the ones used here were selected based on qualitative fit to their findings.

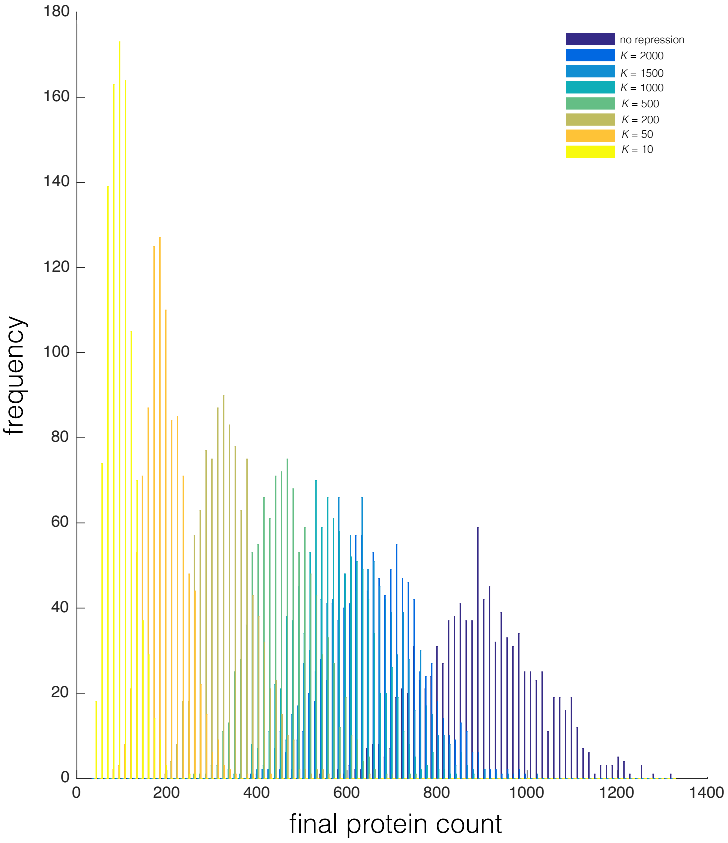

FIGURE 1. Histogram of protein expression levels of eight simulated haploid one-gene autoregulating networks with different dissociation constants. The lower the dissociation constant K, the stronger the autoregulation. Compare with fig. 3a in [19]. Each histogram was obtained from simulating the corresponding network 1000 times and taking the final number of expressed proteins at the end of each simulation run. Each simulation run consisted of simulating gene expression over a simulation time of 10×(protein half-life) seconds, as was done in the original study.

each (i.e. 1/3). The dissociation constants for each pair of genes were also picked randomly from an interval of [1, 1000]. Importantly, all networks simulated in the current study were homozygous, so the dissociation constants were equal across alleles of the same gene.

Table 2. Gene expression noise levels, measured by Fano factor [19], for each of the one-gene autoregulating haploid networks simulated. The smaller the gene dissociation constant, the stronger the autoregulation. The Fano factor is calculated by taking the ratio of the variance to the mean of gene expression, computed over 1000 simulation runs of the given network.

K	Fano factor
no repression	17.43
2000	14.07
1500	13.85
1000	12.15
500	11.26
200	10.86
50	10.71
10	9.75

As opposed to the simulations performed in section 3.1, here we simulated each network for a total simulation time of 20 protein half-lives (equivalent to simulating 20 hours of gene expression, see section 2.3), and assumed the system to have reached equilibrium over the last quarter of this timecourse (i.e. 15 protein half-lives in). Expression level of allele a of gene i at time t was defined as the amount of protein (coded for by that allele) currently present in the system ($p_i^{(a)}$) at time t. Equilibrium expression level was computed by averaging expression level over the equilibrium period. Monoallelic expression was said to occur whenever the difference in equilibrium expression level between the two alleles was at least tenfold.[2]

In our first round of simulations, we drew and simulated 20,000 random networks. We will refer to this group of simulations as experiment 1. Out of a total of 109,982 simulated genes (mean network size was 5.50), 8.57% of genes were classified as monoallelically expressed. Note that any presence of monoallelic expression at all is quite surprising here, as the gene expression model does not include any mechanism for independently regulating the expression of each allele

[2] This difference was computed by dividing the equilibrium expression level of the more expressed allele by the equilibrium expression level of the less expressed allele. If this ratio was greater than 10, the gene was said to be monoallelically expressed. In the case that the less expressed allele had an expression level of 0, it was changed to .0001. This modification, however, introduced the artifact whereby a gene in which one allele is completely silenced (equilibrium expression level of 0) would be classified as monoallelically expressed even when the other allele was only marginally expressed, e.g. at a level of 1 or 2. To avoid such cases, we imposed a lower limit on the expression level of the dominant allele. A gene showing a tenfold difference in allelic expression levels was only classified as monoallelically expressed if the more expressed allele was expressed at a level greater than 2.

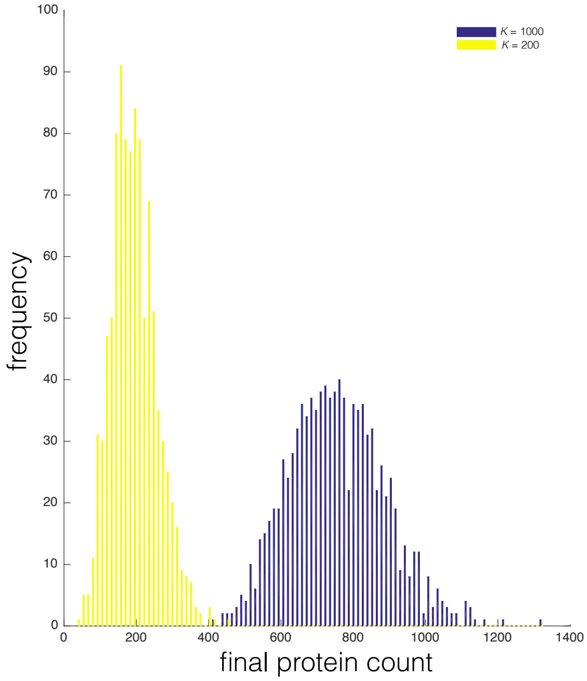

FIGURE 2. Histogram of protein expression levels of a simulated bistable switch network (see text). Each gene has a different dissociation constant. Compare with fig. 4d in Thattai & van Oudenaarden (2001). Histograms obtained as in fig. 1, but for a single two-gene network.

of a gene. The monoallelic expression observed is thus an emergent property of the stochastic system. Strikingly, the amount of monoallelic expression observed in these 20,000 networks resembles that found in the human genome, where ∼10% of autosomal genes show monoallelic expression. [9]

Given that a significant number of genes are being monoallelically expressed, the question to ask is whether any structural properties of the networks these genes are a part of distinguish the genes that are monoallelically expressed from those that aren't. We observed that, compared to their biallelic counterparts, monoallelically expressed genes tended to have the following properties:

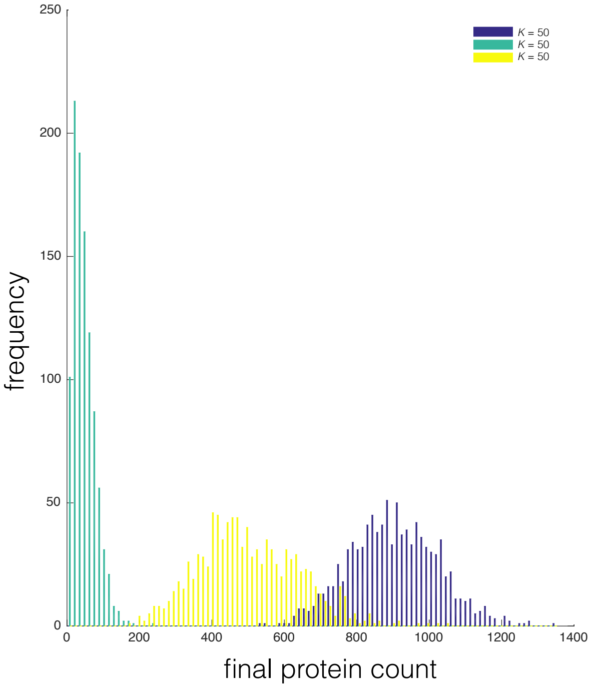

Figure 3. Histogram of protein expression levels of a simulated gene network consisting of three genes in a feed-forward cascade of negative regulation. Compare with fig. 4e in Thattai & van Oudenaarden. Histograms obtained as in fig. 1.

(i) larger network size (i.e. more genes in the network), fig. 4.
(ii) greater network connectivity, fig. 5.
(iii) more regulators (particularly, more positive regulators), figs. 6.

We also found that monoallelically expressed genes tended to have lower total (summed across both alleles) expression levels and lower total expression variance (fig. 7). One might also ask whether monoallelically expressed genes show lower gene expression noise than their biallelic counterparts, in line with previous findings showing that certain gene network properties lead to reduction in gene expression noise. [19] Comparing the total gene expression Fano factor for each group of genes, we in fact found that the monoallelically expressed genes show lower gene expression noise (fig. 8). However, the difference in Fano factor is 2, which, while statistically significant, is likely insignificant biologically. Under our model of gene expression, monoallelic expression does not seem to contribute to reducing gene expression noise. It does, on the other hand, seem to reduce overall expression levels and fluctuations. Alternatively, the causality may be the other way around: lower and less variable expression may lead to a gene being monoallelically expressed.

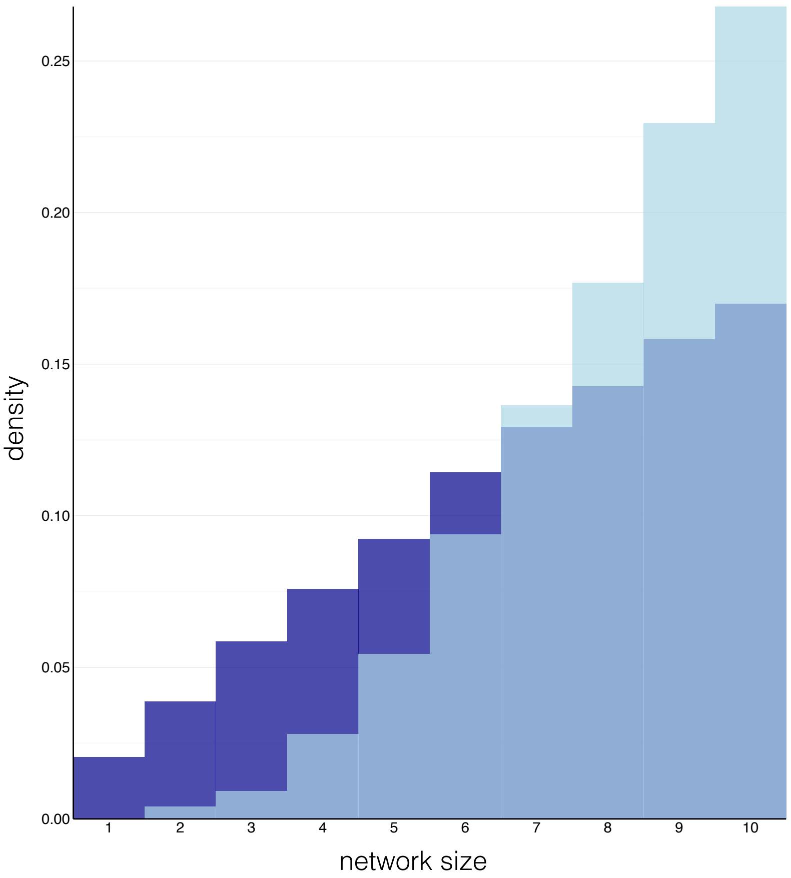

Figure 4. Histogram of network sizes, for monoallelically expressed (MAE) genes (light blue) and biallelically expressed (BAE) genes (dark blue). Density is plotted on the y-axis, rather than absolute frequency, so that the distribution of both populations can be compared. Mean network size was 8.10 and 6.89 for monoallelic and biallelic genes, respectively. A χ^2 test of independence showed the relationship between network size and monoallelic expression is significant ($\chi^2(9) = 2118.7, p < .001$).

These results indeed suggest that the occurrence of monoallelic expression in a given gene depends on the structural properties of the gene expression network it

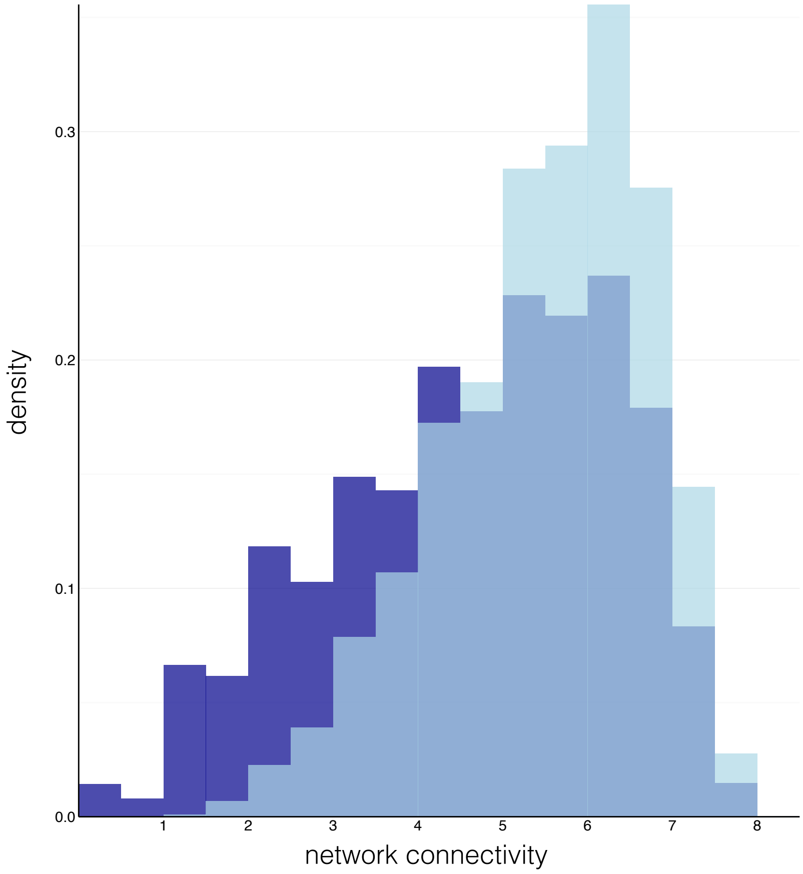

Figure 5. Histogram of network connectivity, for MAE and BAE genes in experiment 1. Network connectivity of a gene is defined as the average number of regulators per gene in the network that it is a part of (all genes in the same network will have the same network connectivity, it is a network-level property). Bin width is 0.5. Mean network connectivity was 5.45 and 4.59 for monoallelic and biallelic genes, respectively; this difference is statistically significant (Mann-Whitney $U = 6.09 \times 10^8, p < .001$, two-tailed test).

is a part of: a gene with many positive regulators that is embedded in a larger and denser network is more likely to undergo monoallelic expression than one with fewer regulators in a smaller and sparser network. To get a measure of the robustness of this dependency, we re-simulated genes 100 times to see if occurence of monoallelic expression was consistent across simulation runs of the same gene. Upon re-simulating 10 genes with prominent monoallelic expression patterns in experiment 1, only two replicated their monoallelic expression in more than 10 out of 100 simulation runs, another two in fact being *biallelically* expressed in all 100 re-simulations. Figure 9 shows the logarithm of the ratio of allele expression

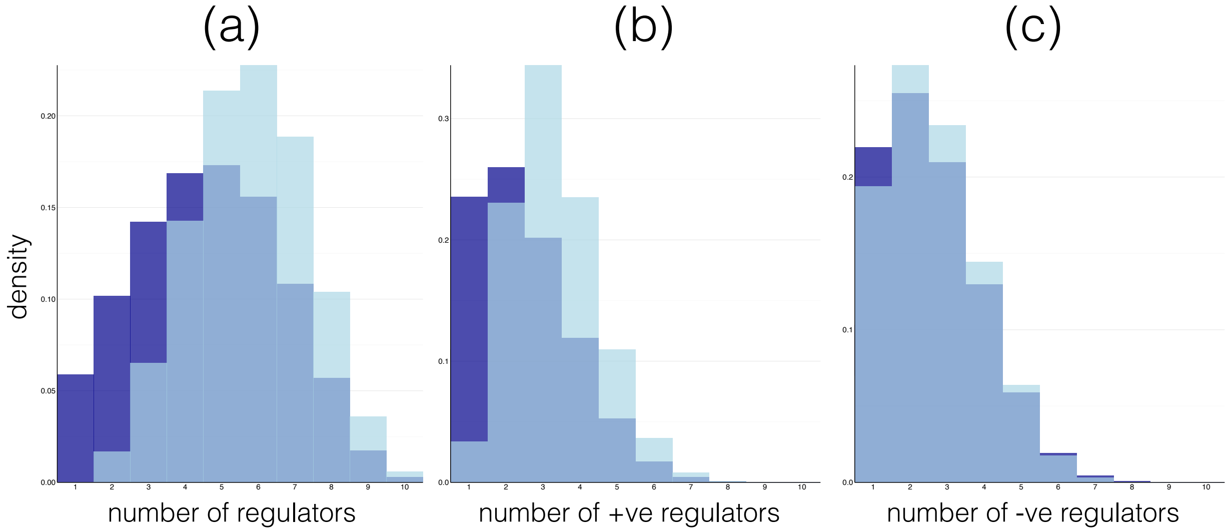

Figure 6. Histograms of (a) total number of regulators, (b) number of positive regulators, and (c) number of negative regulators, for MAE and BAE genes in experiment 1. Mean total number of regulators was 5.77 and 4.56 (positive: 3.30 and 2.24, negative: 2.47 and 2.32) for monoallelic and biallelic genes, respectively. A χ^2 test of independence showed the relationship between number of regulators and monoallelic expression is significant for all regulators ($\chi^2(9) = 3089.9, p < .001$), positive regulators ($\chi^2(9) = 5043.8, p < .001$), and negative regulators ($\chi^2(9) = 183.36, p < .001$).

(more expressed allele:less expressed allele) for each of the 100 simulation runs of one of the re-simulated genes. This gene showed monoallelic expression in only nine of the 100 simulations, despite showing a prominent monoallelic behavior in its original simulation in experiment 1.

These results suggest than monoallelic expression in our simulated genes is not at all robust to the stochasticity of our gene expression model. Crucially, it implies that when a simulated gene shows monoallelic expression, this does not guarantee that that gene is particularly susceptible to being monoallelically expressed - the monoallelism shown in this instance may simply be a rare event. To filter out such rare events and obtain a more sensitive measure of susceptibility to monoallelic expression, we simulated each network 100 times and measured the proportion of simulation runs in which the equilibrium expression levels demonstrated monoallelic expression. We term this measure the *probability of monoallelic expression* of a gene.

Following the above procedure, we drew and simulated 200 random networks, this time simulating each one 100 times to obtain the probability of monoallelic expression of each gene (experiment 2). In this case, a great majority of genes (74%) showed a non-zero probability of monoallelic expression. Restricting the

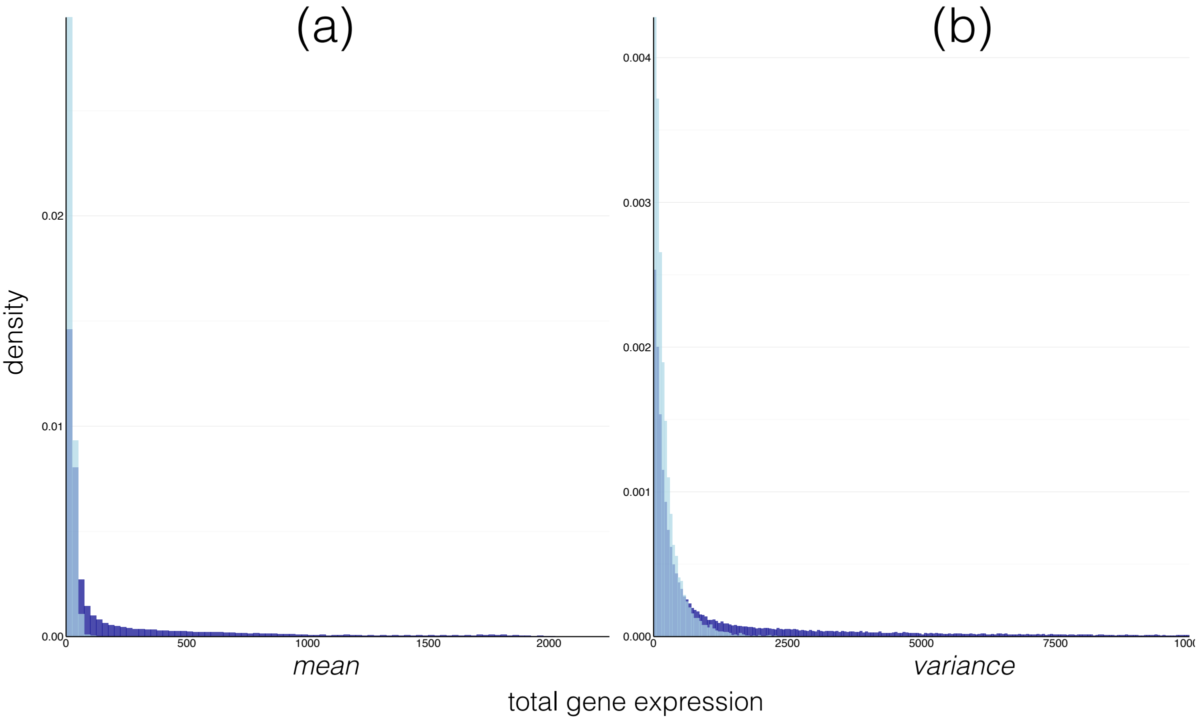

Figure 7. Histogram of (a) equilibrium gene expression mean and (b) equilibrium gene expression variance for MAE and BAE genes in experiment 1. Gene expression is summed across both alleles. Bin width is 25 for mean and 50 for variance. Mean equilibrium gene expression was 19.87 for monoallelic genes and 223. 79 for biallelic genes (although median was 16.9 and 37.7, respectively; this might actually be a better measure due to large spread of expression levels). Mean equilibrium gene expression variance was 238.68 and 3189.07 for monoallelic and biallelic genes, respectively (median: 134, 398). Each of these differences was statistically significant (Mann-Whitney $U = 2.43 \times 10^8, p < .001$, two-tailed test for expression mean; Mann-Whitney $U = 2.84 \times 10^8, p < .001$, two-tailed test for expression variance).

class of monoallelically expressed genes to those with greater than 20% chance of being monoallelically expressed, 17.6% of the 1073 simulated genes can be said to show monoallelism. This number falls within the 10-20% range proposed by Savova *et al.* (2013) for the proportion of monoallelically expressed genes in the human genome, extrapolating from the limited number of cell types in which monoallelism has been studied. [17]

Using this new measure of monoallelic expression, we can go back and verify the observations made in experiment 1 (figs. 10,11,12). These were confirmed statistically with linear regression. Having observed that regulatory connections and low expression variance are characteristic of monoallelically expressed genes,

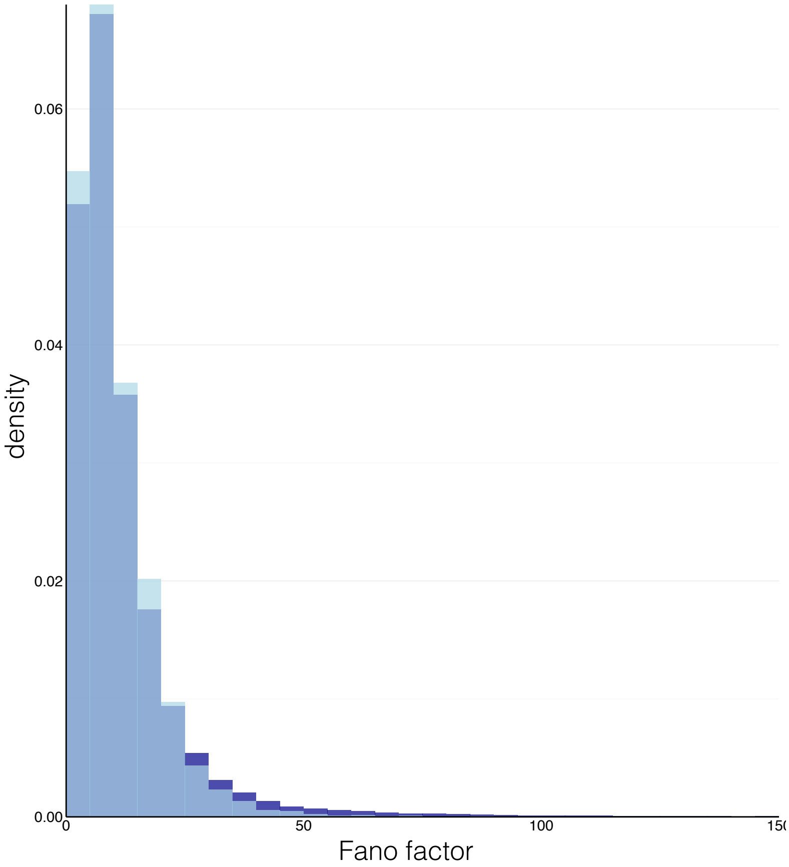

FIGURE 8. Histogram of Fano noise for MAE and BAE genes in experiment 1. Fano noise is computed by taking the ratio of the expression variance to the expression mean (over the equilibrium period). Bin width is 5. Mean Fano noise was 12.30 and 10.10 for monoallelic and biallelic genes, respectively; this difference is statistically significant (Mann-Whitney $U = 4.55 \times 10^8, p < .001$, two-tailed test).

we also asked whether autoregulation would be as well (as autoregulation can lead to reduced expression noise $\propto$ expression variance [19, 18]). Comparing the average probability of monoallelic expression for genes with negative autoregulation, positive autoregulation, and no autoregulation, we found that genes with positive autoregulation were in fact about twice as likely to be monoallelically expressed (fig. 13) Overall, our findings from experiment 2 confirm the findings from experiment 1: more regulatory interactions in a gene's network and more positive regulation of that gene lead to increased susceptibility to monoallelic expression. Also, genes more likely to be monoallelically expressed are expressed at a lower level, with smaller temporal fluctuations (fig. 14).

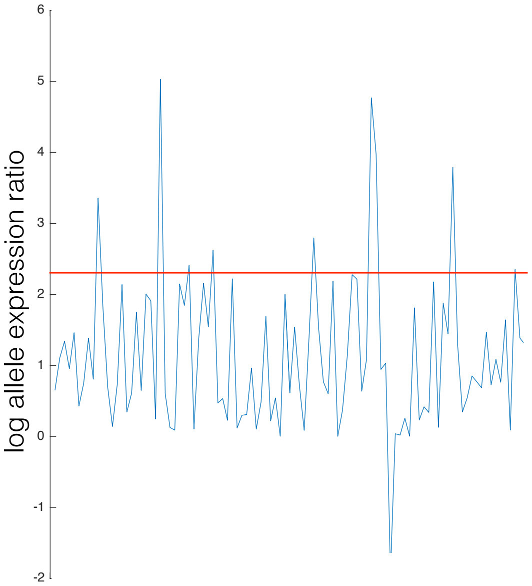

Figure 9. Natural logarithm of allele expression ratio across simulation runs of a gene shown to be monoallelically expressed in experiment 1. Allele expression ratio is the ratio of the equilibrium expression level of the more expressed allele to the equilibrium expression level of the less expressed allele. Red line indicates threshold above which a gene is classified as being monoallelically expressed (ratio > 10). All parameters are exactly the same for each simulation run, the x-axis simply represents 100 simulations of the same network. Note the stochasticity of allele expression levels: this gene is monoallelically expressed in only nine of the 100 simulation runs, and its allele expression ratio fluctuates dramatically from simulation to simulation.

3.3. **Experiments 3 & 4: assessing the effect of number of regulators on probability of monoallelic expression.** It is important to note that the number of regulators of a given gene is not independent of its network size. For example, a gene that has 8 positive regulators must be part of a network with at least 8 genes in it. Furthermore, because that gene is in a larger network, it is more likely to also have negative regulators. To disentangle the effects of network size from the effect of number of regulators on probability of monoallelic expression, we simulated 300 ten-gene networks with random connections and dissociation constants (experiment 3). We again found the same patterns (fig. 15), confirming that, keeping network size constant, increasing the number of positive regulators increases the probability of monoallelic expression. Moreover, the regression for number of negative regulators in this case showed a highly significant negative slope, suggesting that increasing the number of negative regulators may in fact decrease the probability of expression. While the fit of the regression to the data

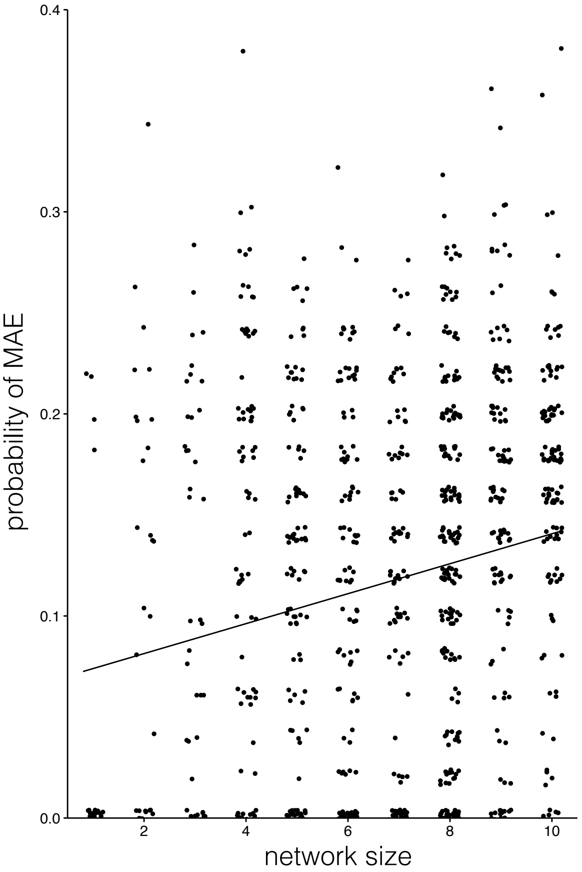

FIGURE 10. For each simulated gene in experiment 2, the size of its network plotted against the probability of monoallelic exression (MAE). This was computed by taking the proportion of simulation runs out of 100 in which the gene showed monoallelic expression in the equilibrium period. To improve visibility, each point is randomly jittered along both dimensions. Line represents linear regression showing a statistically significant relationship between the two variables ($r^2 = .0349, p < .001$). I should note here that linear regression isn?t quite appropriate for this data since the data points are not independent of each other (e.g. genes from the same network will have the same network size). However, the effect is clear and captured by the linear regression line, so I chose to use linear regression for the sake of consistency.

is quite low (the regression equation accounts for 5% of the variance), it is worth noting that it is 10 times higher than in experiment 2 (and the p value is lower).

We can continue to probe the effect of regulatory connections on probability of monoallelic expression by manipulating the number of connections in a given gene network. To do this, we picked a gene network that contained genes with high

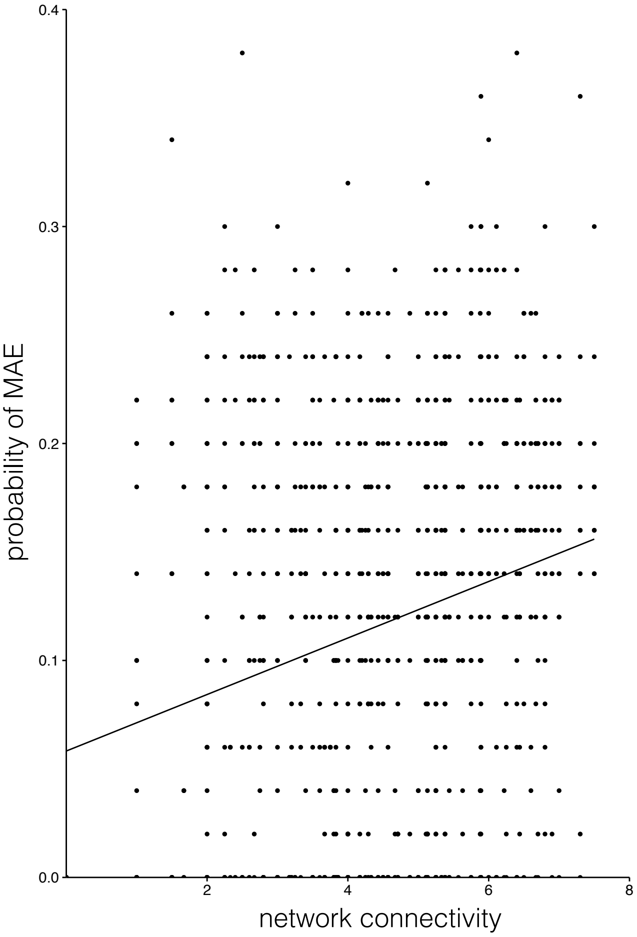

FIGURE 11. For each simulated gene in experiment 2, the connectivity of its network as defined above (see fig. 5) plotted against the probability of MAE. Line represents linear regression showing a statistically significant relationship between the two variables ($r^2 = .0516, p < .001$).

probability of monoallelic expression. We then generated variants of this network with random samples of its connections. By sampling different numbers of connections, we generated five groups of variants with different numbers of connections (experiment 4). We then averaged the probability of monoallelic expression over genes in each group. As expected, we found that increasing the number of connections in the network increased mean probability of monoallelic expression of its genes (fig. 16). In fact, no genes in any of the variants with 9 connections (the minimum tested) showed monoallelic expression in any of their simulations.

3.4. **Experiment 5: assessing the temporal stability of monoallelic expression.** Given the above highlighted stochasticity of monoallelic expression in

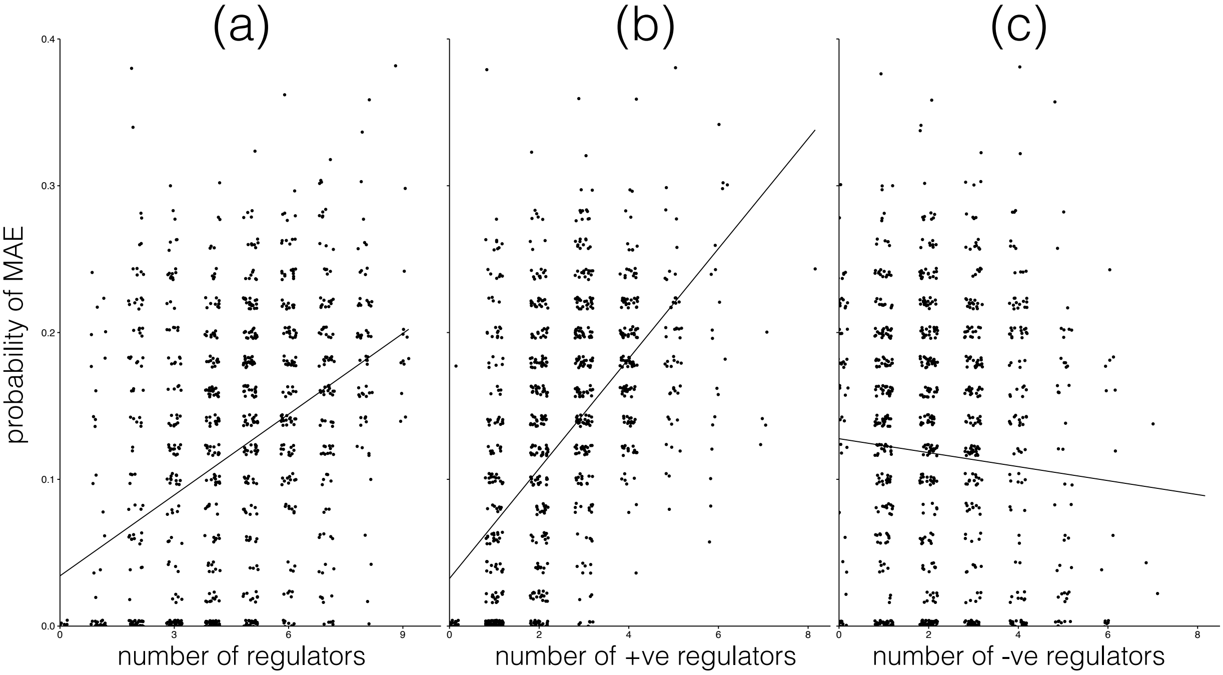

FIGURE 12. For each simulated gene in experiment 2, (a) total number of regulators, (b) number of positive regulators, and (c) number of negative regulators, plotted against the probability of MAE. To improve visibility, each point is randomly jittered along both dimensions in all plots. Lines represent linear regression showing a statistically significant relationship between the probability of MAE and total number of regulators ($r^2 = .156, p < .001$), number of positive regulators ($r^2 = .361, p < .001$), and number of negative regulators ($r^2 = .00551, p = .015$).

our gene expression model, it is important to ask whether genes shown to be monoallelically expressed during the equilibrium period continue to be so if we extend the simulation time. We thus picked sixty networks from experiment 2 with a range of probabilities of monoallelic expression and quintupled the simulation time (100 protein half-lives, as opposed to 20) (experiment 5). Monoallelic expression was determined by taking the ratio of each allele's mean expression level over the entire extended equilibrium period. The starting point of the equilibrium was thus kept the same as in the shorter timecourses at 15 protein half-lives in, making the equilibrium period 17 times longer than in experiments 1-4 (85 protein half-lives vs 5).

As might be expected from the observed volatility of monoallelic expression patterns, monoallelism was never maintained over this extended equilibrium period. By comparing the monoallelic expression present in the short experiment 2-sized equilibrium period to that in the extended equilibrium period for each gene, we can

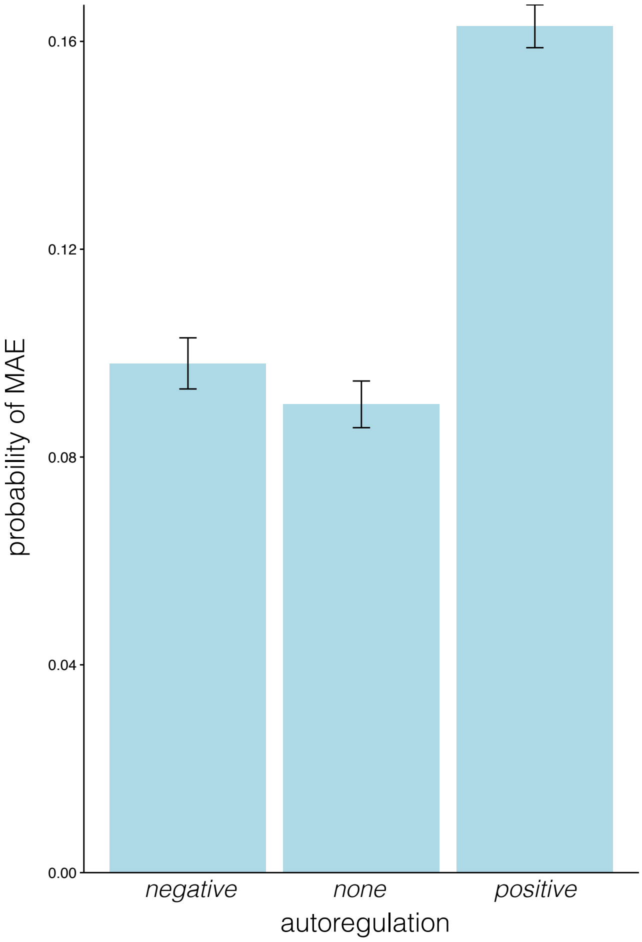

FIGURE 13. Mean probability of MAE of genes in experiment 2, grouped by autoregulation (positively autoregulated, negatively autoregulated, or not autoregulated). Error bars reflect standard error of the mean. Mean probability of MAE was significantly higher for genes with positive autoregulation (M = .163, SD = .0779) than for those with negative autoregulation (M = .0980, SD = .0938) or no autoregulation (M = .0901, SD = .0844). This difference was statistically significant (Kruskal-Wallis $\chi^2(2) = 136.1, p < .001$).

immediately see that, whereas monoallelism can be observed to occur relatively often within the shorter equilibrium period, it in fact *never* lasts long enough to be said to occur over a longer timescale (fig. 17).

To further examine the temporal stability of monoallelic expression, we took each gene expression timecourse and computed the average number of timepoints said to lie within a period of stable monoallelic expression (defined as monoallelic expression over at least 500 timepoints). The result was 2,282.1. Multiplying this

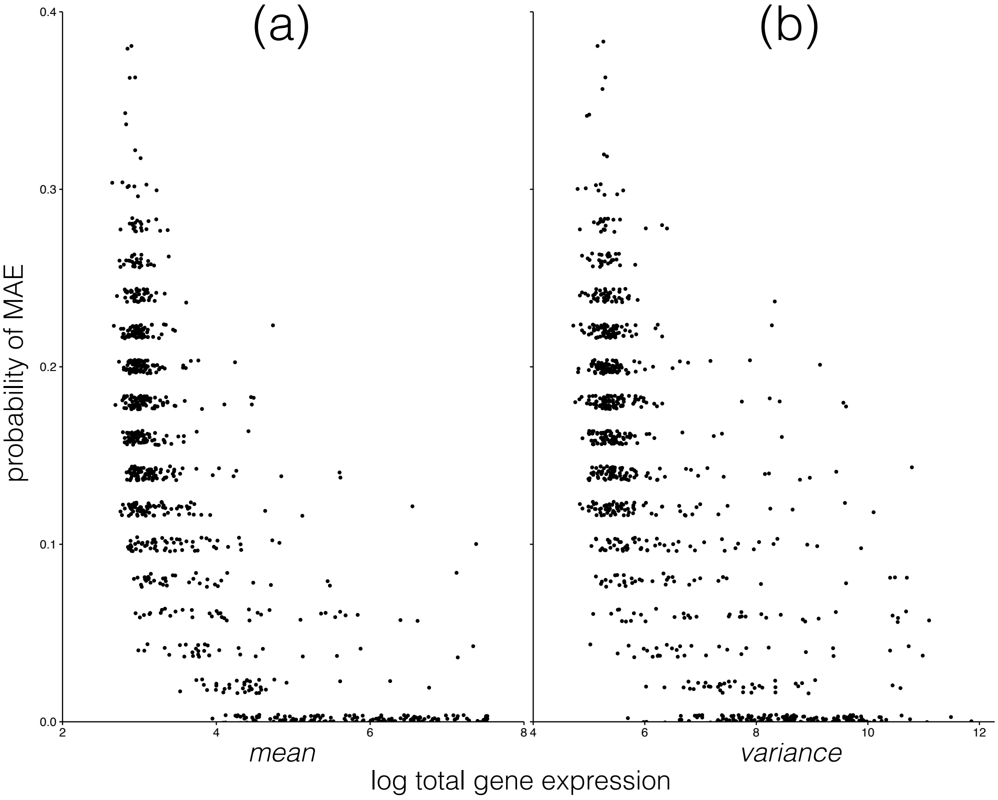

FIGURE 14. For each simulated gene in experiment 2, (a) equilibrium expression mean and (b) equilibrium expression variance plotted against the probability of MAE. Log expression is plotted so as to include the full range. To improve visibility, each point is randomly jittered along the y-axis only. Line represents linear regression showing a statistically significant relationship between the two variables. Note that almost all genes with log expression mean and variance below ~5 have probability of MAE > 0.

by the mean simulation time between timepoints, this comes out to 500 seconds $\approx$ 10 minutes of gene expression time. For rapidly proliferating eukaryotes like *S. cerevisiae* with a doubling time of ~90-120 minutes [4], this is a marginally biologically significant amount of time. The distribution of average time spent in stable monoallelic expression for each gene is plotted in figure 18. These periods of stable monoallelic expression varied greatly in their duration and in the identity of which allele was being expressed (fig. 19).

Taking advantage of the fact that the networks used in this experiment were already simulated in experiment 2, we can use the timecourses shortened to the original simulation time to check the robustness of our probability measure. For each gene, we compared the probability of monoallelic expression observed in this

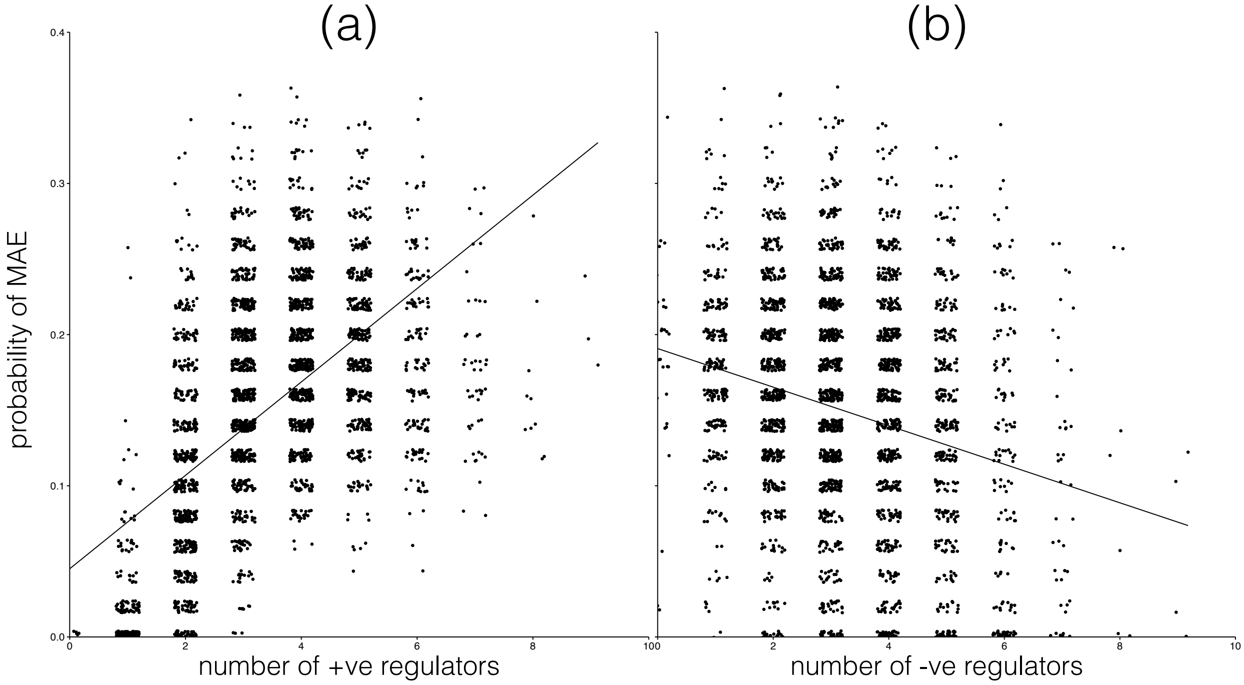

Figure 15. For each simulated gene in experiment 3, (a) number of positive regulators and (b) number of negative regulators plotted against the probability of MAE. To improve visibility, each point is randomly jittered along both dimensions in each plot. Lines represent linear regression showing a statistically significant relationship between probability of MAE and number of positive regulators ($r^2 = .325, p < .001$), and between probability of MAE and number of negative regulators ($r^2 = .0545, p < .001$).

experiment - determined using the experiment 2-sized equilibrium period - to that observed in experiment 2 (fig. 20). The probabilities were highly correlated ($r = 0.728$), indicating that our measure of probability of monoallelic expression is relatively robust to the stochasticity of our gene expression model.

4. Discussion

We asked whether certain gene network properties could lead to monoallelic expression. By drawing random networks and simulating them, we found that monoallelically expressed genes tended to share certain properties:

(1) the regulatory networks they are a part of contain many genes and regulatory connections between them
(2) they are highly positively regulated, and also negatively regulated but to a much lesser extent
(3) on average, they are expressed at low levels, with low variance

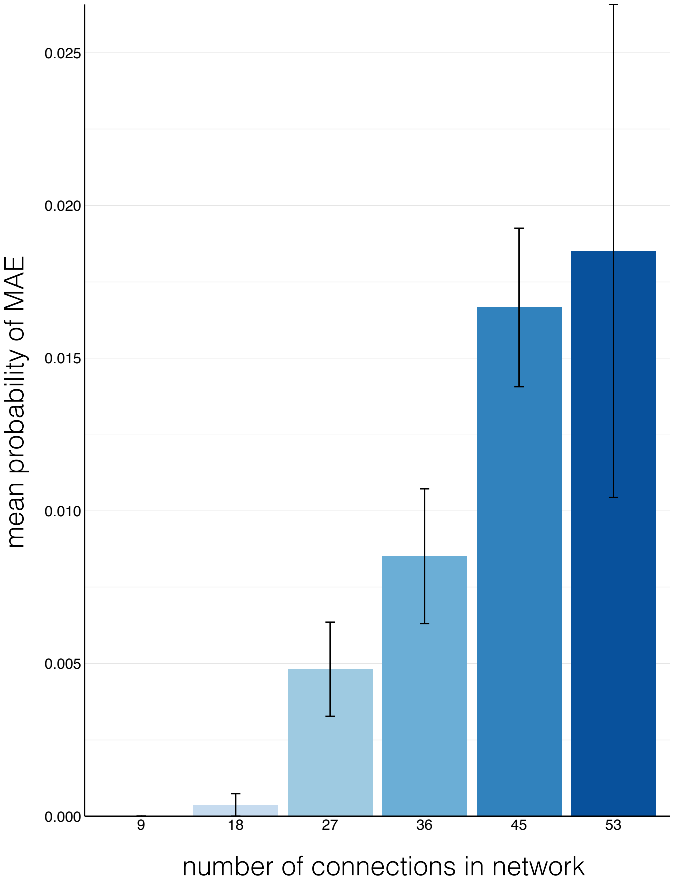

FIGURE 16. Mean probability of monoallelic expression of genes in each group of network variants. Error bars represent standard error of the mean. Each network variant was generated by removing a random sample of the original network's connections. In this manner, variants of the original network with 9, 18, 27, 36, or 45 connections were constructed. Aside from the number of connections, all other aspects of these networks were the same as the original network (dissociation constants, size, etc.). These variants were then grouped by number of connections, and averages were taken over all genes in all the networks in each group. The rightmost bar corresponds to the original network, which had 9 genes and 53 connections. The standard error for this measurement is substantially larger because genes were overaged over only one network, whereas there were 10 network variants in each of the other groups. Differences were statistically significant (Kruskal-Wallis $\chi^2(5) = 76.47, p < .001$).

It is worth noting a few quantitative observations from our simulations that underscore these tentative conclusions. Firstly, a gene must be positively regulated to be monoallelically expressed - with very few exceptions, genes with no positive

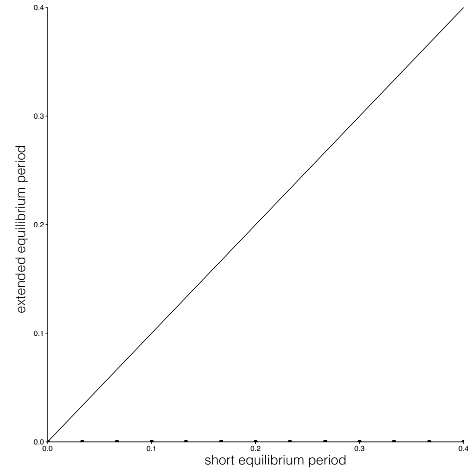

FIGURE 17. For each gene simulated in experiment 5, probability of MAE measured in short equilibrium period plotted against probability measured in the extended equilibrium period. Short equilibrium period duration was equated to that used in experiment 2 (5 protein half-lives. Extended equilibrium period was 17 times longer (85 protein half-lives, see text). The spread of all the data points along the x-axis reflects the fact that MAE was never maintained across the extended equilibrium period. Plotted line is $y = x$.

regulation were not monoallelically expressed. Secondly, more than seven negative regulators leads to very low probability of monoallelic expression, although an intermediate amount seems to be more effective than none at all in eliciting monoallelism. Finally, every single gene with equilibrium expression below ~60 had a non-zero probability of expression. The same goes for variance of equilibrium expression below 1000. Virtually no genes with expression level and variance above these values had probability of monoallelic expression above 20%.

The main finding here, however, is that these properties are not determinant of whether monoallelic expression occurs: they only influence the probability that a given gene be monoallelically expressed. Our simulations demonstrate that under highly stochastic gene expression, monoallelic expression can emerge, albeit stochastically as well. This is quite a surprising finding, given that our stochastic gene expression model did not incorporate any mechanism for indpendently regulating the expression of each allele of a gene. The monoallelic expression observed was highly stochastic in that (i) it is rarely maintained for more than 10 minutes at a time (in units of gene expression time), and (ii) the identity of the expressed/silenced gene can change from one period of monoallelic expression to another within the same expression timecourse of a gene.

One interpretation of these findings is that, in our mathematical model of gene expression, expression of each allele is a Markov random walk, where the state

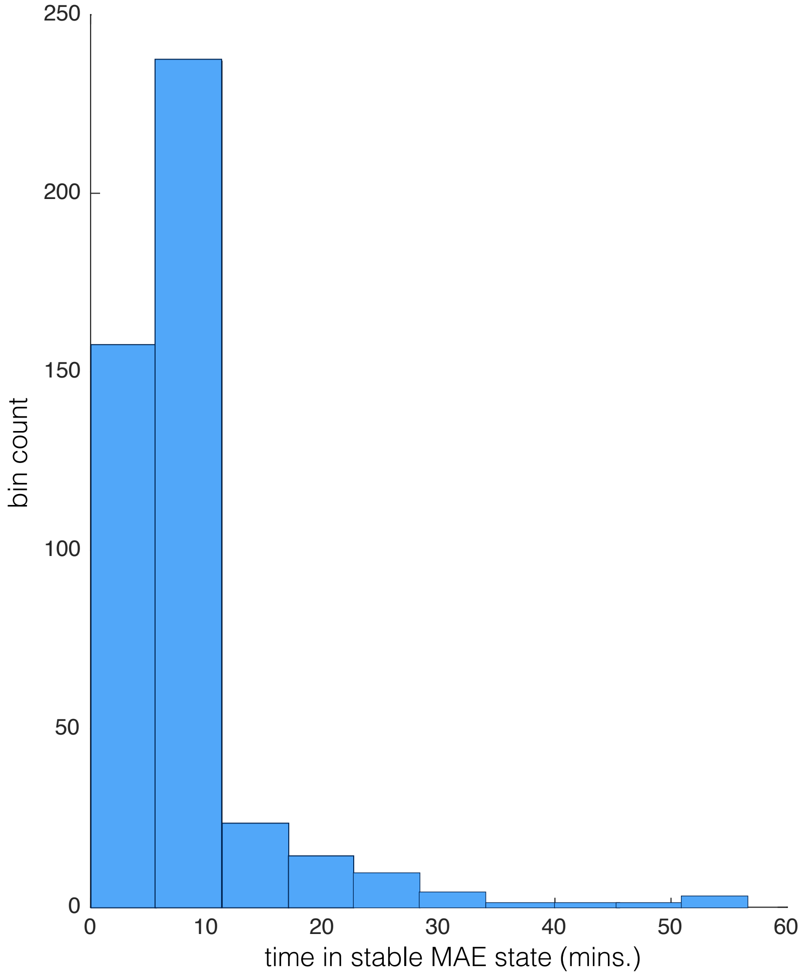

FIGURE 18. Histogram of gene expression time spent in a stable period of monoallelic expression (see text). These were computed for each simulated gene in experiment 5, over full extended timecourse.

of the system at a given timepoint is random and solely dependent on the state of the system at the immediately preceding timepoint. [7] The gene properties outlined above could be acting to ensure the state of the system remain within a constrained space within which the probability distribution over subsequent states is skewed toward states of monoallelism. Particularly, positive feedback of gene expression seems critical for keeping the system in this space.

This hypothesis is testable. If it is true, it should be possible to formally derive the moments of the probability distribution over states to estimate the expected frequency of monoallelic expression. This is a natural next step in the project, whereby we could compare the expected frequency under this hypothesis to the

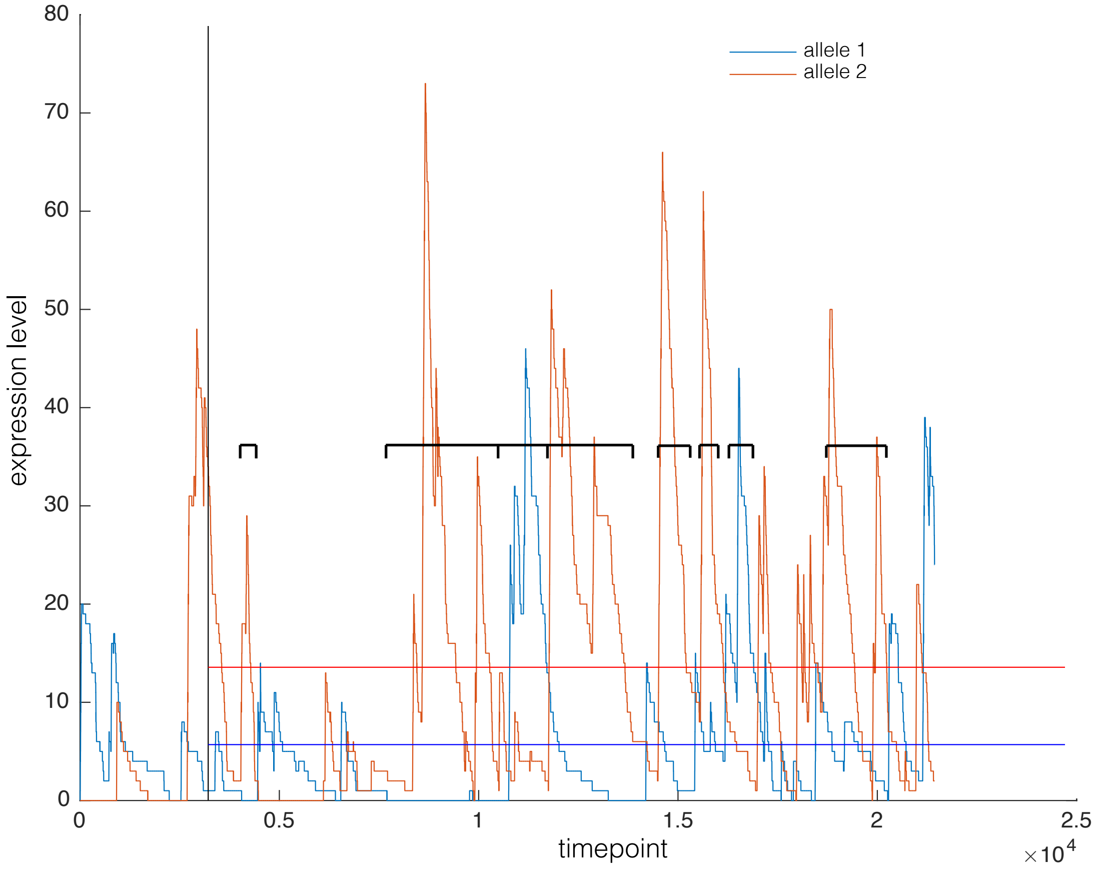

Figure 19. Timecourse of a simulated gene from experiment 5, demonstrating the instability of monoallelic expression. Horizontal brackets indicate periods of stable monoallelic expression. Note that the identity of the expressed/silenced allele changes from period to period, as do their lengths. Horizontal lines indicate mean allele expression levels over the equilibrium period, which begins at the black vertical line. Note that x-axis is not time. Each data point indicates the expression level of an allele of the gene at a timepoint in the timecourse. Each timepoint corresponds to an instance in which some reaction occurred in the gene expression simulation (transcription/translation/mRNA degradation/protein degradation).

observed frequency from our simulations (∼8% of random networks show monoallelic expression in a given instance).

The implications of such a finding would be significant for thinking about diploid evolution of heterozygous individuals. Evolutionary theorists often model evolution in a population of heterozygotes by taking the fitness of a given individual to be the fitness provided by the phenotype expressed by the dominant allele, or

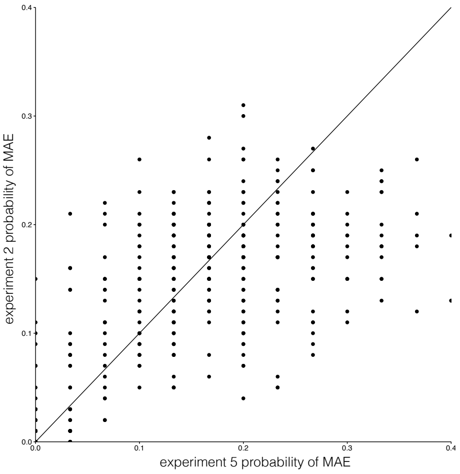

Figure 20. For each gene simulated in experiment 5, probability of MAE observed in short equilibrium period plotted against the probability of MAE measured in experiment 2 (where the same genes were simulated but for a shorter simulation time, see text). Probabilities for each gene were correlated ($r = .728$), indicating that probability of MAE is a robust property of genes simulated under our gene expression model. Plotted line is $y = x$.

both alleles in the case of codominance. [11] If you take into account that there is a non-zero probability of monoallelic expression of either allele, this no longer becomes correct. If monoallelic expression of a given gene occurs with, say, 10% probability, then 10% of the heterozygote individuals will express the phenotype of only one of their two alleles, irrespective of dominance. To accurately model evolution in this population, these individuals' corresponding fitness scores would then have to be changed accordingly. If our preliminary results are correct that about 15% of genes have ∼20% probability of being monoallelically expressed, this is an important factor to consider.

In this way, monoallelic expression can also accelerate genetic drift by increasing positive selection for genetic mutations. [1] When a new mutation is introduced into a population of diploids, it appears in a heterozygous individual. In the case of biallelic expression, the survival advantages of the phenotype expressed by this mutation might be masked or dampened by the expression of the other allele. On the other hand, monoallelic expression would allow for the phenotype to be expressed in all its fullness, providing the individual with the mutation the accompanying survival benefits and thus aiding positive selection for it. However, monoallelic expression can also relax negative selection for deleterious mutations in the same way, by masking their corresponding undesirable phenotypes. It is easy to see how probabilistic monoallelic expression as that predicted by our model could be advantageous for a diploid population.

In considering how this stochastic interpretation of monoallelic expression holds in the context of current empirical investigations of monoallelic expression, several issues immediately arise. Firstly, the stochastic random walk model of monoallelic expression cannot account for the stability and inheritance of monoallelic expression found in actual cells. [9, 10] While it does correctly predict that a given gene will be monoallelically expressed in some cells but not in others [6, 12], when a cell expresses only one allele of a gene, it maintains this pattern of expression over time and across mitotic divisions. The identity of the expressed does not change. While these findings directly contradict the behaviour of our model, some otherwise necessary modifications to our gene expression model might allow it to account for them. A consistent finding across monoallelically expressed genes is that their alleles show differential epigenetic marking, in terms of chromatin [15], DNA methylation [13], and histone modifications [6]. A natural next step to improving our model is to incorporate chromatin states. The dynamics of this more complete system could allow for the stochastic monoallelic expression arising from the gene-level regulatory dynamics to be stabilized over longer timescales, and to be clonally inherited.

Other findings in the literature agree with the hypothesis that monoallelism could arise from stochastic gene expression. Indeed, it has been proposed that monoallelic expression arises via a stochastic mechanism [6, 13, 12], and argued that such a mechanism would be advantageous in certain cases. [12] Our finding that monoallelically expressed genes tend to show lower expression levels is also supported by the literature. For example, Gimelbrant *et al.* (2007) found that clones expressing *APP* monoallelically had a lower level of expression than their biallelic counterparts, and Eckersley-Maslin *et al.* (2014) found lower median expression level of monoallelically than biallelically expressed genes in neural progenitor cells. Using a classifier trained on chromatin signatures of monoallelism to identify monoallelically expressed genes, Nag *et al.* (2013) found that those genes showed lower expression levels than the genes classified as biallelic. Moreover, the distribution they found for each group of genes resembled that found here (fig. 21).

Should the hypothesis that monoallelic expression is elicited by certain configurations of gene networks be confirmed upon subsequent elaboration of our gene expression model, an empirical test should be sought. One possibility here is to search for gene networks with those configurations, and asking whether the genes in those networks are monoallelically expressed. A second possibility is to construct artificial gene networks and see how they behave with respect to monoallelism.

5. Conclusions

We found that MAE can arise from a stochastic gene expression model with no mechanism built in to independently regulate alleles. We hypothesize that in fact this process is a Markov random walk, and that certain gene properties

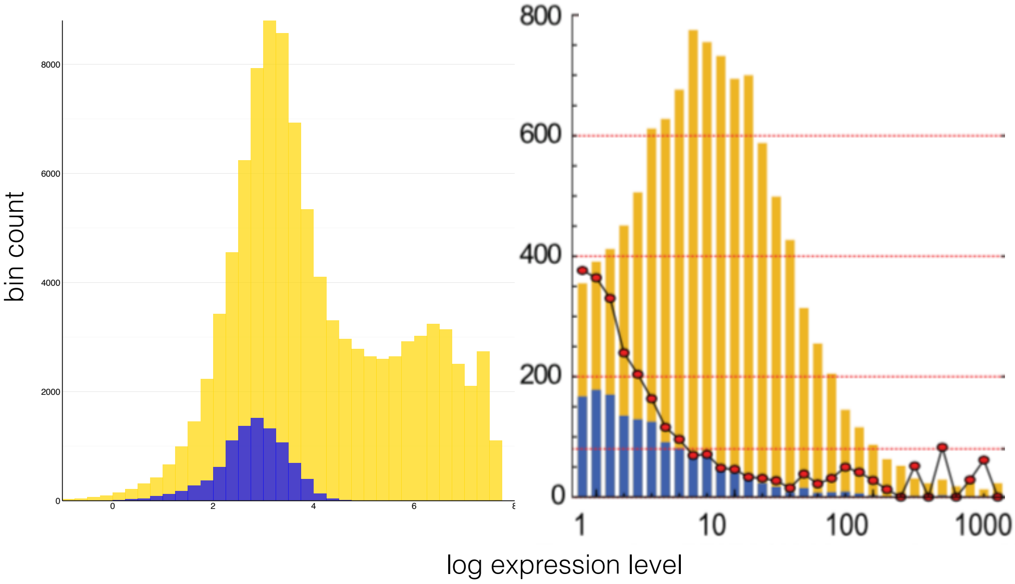

Figure 21. Histogram of gene expression level for monoallelically expressed genes (blue) and biallelically expressed genes (gold). On the left, figure taken from Nag *et al.* (2013). These genes were classified as monoallelic or biallelic based on a decision-tree classifier trained on chromatin signatures of monoallelic expression. On the right, histogram generated from data from experiment 1.

will increase the probability of the next step in the random walk to lead towards monoallelic expression. Although purely stochastic and incorporating only gene regulation mechanisms, this process can give rise to monoallelic expression over marginally biologically significant timescales (∼10 minutes).

We propose that this model be analyzed further in light of the Markov random walk hypothesis. Additionally, it should be augmented to incorporate chromatin states, which would allow it to better reproduce empirical findings regarding the stability and heritability of monoallelic expression.

References

[1] Chess, A. Mechanisms and consequences of widespread random monoallelic expression. *Nature Reviews Genetics 13*, 6 (2012), 421–428.

[2] Chess, A. Random and non-random monoallelic expression. *Neuropsychopharmacology 38*, 1 (2013), 55–61.

[3] Chess, A., Simon, I., Cedar, H., and Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell 78*, 5 (1994), 823–834.

[4] Cooper, G. M. *The Cell: A Molecular Approach, 2nd edn.* Sunderland, MA: Sinauer Associates, 2000.

[5] Eckersley-Maslin, M. A., and Spector, D. L. Random monoallelic expression: regulating gene expression one allele at a time. *Trends in Genetics 30*, 6 (2014), 237–244.

[6] Eckersley-Maslin, M. A., Thybert, D., Bergmann, J. H., Marioni, J. C., Flicek, P., and Spector, D. L. Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Developmental Cell 28*, 4 (2014), 351–365.

[7] Gardiner, C. W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer-Verlag, 1983.

[8] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry 81*, 25 (1977), 2340–2361.

[9] Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., and Chess, A. Widespread monoallelic expression on human autosomes. *Science 318*, 5853 (2007), 1136–1140.

[10] Gimelbrant, A. A., Ensminger, A. W., Qi, P., Zucker, J., and Chess, A. Monoallelic expression and asynchronous replication of p120 catenin in mouse and human cells. *Journal of Biological Chemistry 280*, 2 (2005), 1354–1359.

[11] Gokhale, C. S., and Traulsen, A. Evolutionary multiplayer games. *Dynamic Games and Applications 4*, 4 (2014), 468–488.

[12] Guo, L., Hu-Li, J., and Paul, W. E. Probabilistic regulation in th2 cells accounts for monoallelic expression of il-4 and il-13. *Immunity 23*, 1 (2005), 89–99.

[13] Jeffries, A. R., Perfect, L. W., Ledderose, J., Schalkwyk, L. C., Bray, N. J., Mill, J., and Price, J. Stochastic choice of allelic expression in human neural stem cells. *Stem cells 30*, 9 (2012), 1938–1947.

[14] Keverne, B. Monoallelic gene expression and mammalian evolution. *Bioessays 31*, 12 (2009), 1318–1326.

[15] Nag, A., Savova, V., Fung, H.-L., Miron, A., Yuan, G.-C., Zhang, K., and Gimelbrant, A. A. Chromatin signature of widespread monoallelic expression. *Elife 2* (2013), e01256.

[16] Pernis, B., Chiappino, G., Kelus, A. S., and Gell, P. G. Cellular localization of immunoglobulins with different allotypic specificities in rabbit lymphoid tissues. *The Journal of experimental medicine 122*, 5 (1965), 853–876.

[17] Savova, V., Vigneau, S., and Gimelbrant, A. A. Autosomal monoallelic expression: genetics of epigenetic diversity? *Current opinion in genetics & development 23*, 6 (2013), 642–648.

[18] Stewart, A. J., Seymour, R. M., Pomiankowski, A., and Reuter, M. Underdominance constrains the evolution of negative autoregulation in diploids. *PLoS Comput Biol 9*, 3 (2013), e1002992.

[19] Thattai, M., and Van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences 98*, 15 (2001), 8614–8619.