

Research article

Transcriptional coupling of neighbouring genes and gene expression noise: evidence that gene orientation and non-coding transcripts are modulators of noise.

Guang-Zhong Wang^{1,2}, Martin J. Lercher² and Laurence D. Hurst¹

1. Dept. of Biology and Biochemistry, University of Bath, Bath, UK BA2 7AY
2. Heinrich-Heine-University, Universitaetsstr. 1, D-40225, Duesseldorf, Germany

Contact: l.d.hurst@bath.ac.uk

Running title: Gene orientation and non-coding transcripts as noise modulators

For some genes, notably essential genes, expression when expression is needed is vital hence low noise in expression is favourable. For others noise is necessary for coping with stochasticity or for providing dice-like mechanisms to control cell fate. But how is noise in gene expression modulated? We hypothesise that gene orientation may be crucial, as for divergently organized gene pairs expression of one gene could affect chromatin of a neighbour thereby reducing noise. Transcription of antisense non-coding RNA from a shared promoter is similarly argued to be a noise-reduction mechanism. Stochastic simulation models confirm the expectation. The model correctly predicts: that protein coding genes with bi-promoter architecture, including those with a ncRNA

partner, have lower noise than other genes; divergent gene pairs uniquely have correlated expression noise; distance between promoters predicts noise; ncRNA divergent transcripts are associated with genes that *a priori* would be under selection for low noise; essential genes reside in divergent orientation more than expected; bi-promoter pairs are rare subtelomerically, cluster together and are enriched in essential gene clusters. We conclude that gene orientation and transcription of ncRNAs, even if unstable, are candidate modulators of noise levels.

Abbreviations: CUTs, cryptic unstable transcripts; FOP, optimal codon usage; ncRNA, non-coding RNA; NFR, nucleosome free region; SUTs, stable annotated transcripts; TF, transcription factor; TSS, transcription start site

Introduction

Between genetically identical cells we see variation in abundance of any given

transcript or protein. This variation is noise in gene expression . There is also

considerable variation between genes in the level of noise . In part the between-gene

variation in noise, assayed as the coefficient of variation (standard deviation/mean

across individuals), is accounted for by expression level, there being lower noise for more highly expressed genes . Even controlling for this, using an abundance corrected noise measure, there remains, however, striking variation . What are the underlying determinants of this abundance-independent variation in noise levels between genes and might the variation between genes in their noise levels reflect the activity of selection?

For some genes high noise is likely to be significantly deleterious. In particular, essential genes are, by definition, genes for which reductions (but not necessarily increases) in dosage are highly deleterious. Stochastic fluctuation in abundance of such proteins is thus likely to be highly deleterious as dose can, by chance, sink to fitness-reducing low levels . We should then expect such proteins to be under selection to have

low noise. That they do have low noise is consistent with such a model . Haplo-
insufficient genes have yet lower noise, as might be expected . Conversely noise can be
advantageous to some degree. Noise, for example, can provide the underlying basis of
dice-like behaviour necessary for alternative cell fate specification in a genetically
uniform population of cells (e.g. the developing embryo) . Further, if the environment is
stochastic, noisy gene expression can be an effective mechanism to cope with
uncertainty . Noise in the expression of metabolic import channels is, for example,
potentially advantageous when nutrient availability is fluctuating. It is striking that of
all metabolic genes, import channels are the most noisy . Stress response genes are also
expected to be high noise genes, these also being responsive to an uncertain
environment .

While noise may then be an important target of selection, this leaves the issue of how mechanistically noise is modulated. At the transcript level slow translation rates and low mRNA half lives are likely to reduce noise . Much noise modulation is probably achieved at the transcriptional control level. TATA controlled genes, in particular, tend to be especially noisy and expression noise of genes is increased when the binding site of GAL1 promoter is moved closer to a TATA-box . The underlying cause of an association with TATA is unresolved. The high expression variation of TATA-box containing gene may be owing to the binding stability of transcription-mediating factor TBP or related to the high nucleosome occupancy , suggesting a link to chromatin dynamics. A recent report of the lack of activating histone modifications in this region

supports the latter.

The above results go some way to unifying TATA control with chromatin level control,

also thought to be important in noise modulation . In one striking example , a pair of

genes inserted in tandem showed co-ordinated spiking in their gene expression, while

the same pair when unlinked showed little co-ordination. This result suggests a model

whereby opening of chromatin permits accessibility to transcription factors. Regular

opening and closing of chromatin then leads to co-ordinated expression, and correlated

noise levels, of neighbours. Such a model correctly predicts that across a genome,

controlling for similarity of transcription factor control, linked genes show much higher

levels of co-expression than do unlinked genes . This in turn is related to nucleosome

occupancy . The magnitude of this effect is noteworthy: two random unlinked genes regulated by the same set of transcription factors show no higher co-expression than a pair of linked genes with no similarity in their transcription factors .

This class of model has led to the suggestion that the genomic distribution of essential genes and chromatin control should co-evolve such that essential genes end up clustered into domains with largely open chromatin, thereby ensuring low noise and expression when expression is needed . The model has some predictive power. It correctly predicts, for example, that essential genes should be rare subtelomerically in yeast, these being domains inconsistent with permanently open chromatin. It also correctly predicts nucleosome occupancy in domains rich in essential genes and that noise levels

of non-essential genes should be predicted by the local density of essential genes . Here we extend the logic of chromatin mediated noise modulation to propose that modulation of noise by DNA dynamics might affect gene pairs differentially dependent on their orientation.

Gene pairs can come in one of three orientations: convergent (), co-oriented (or) or divergent (). These three classes are not equally conserved. In human, mouse, and rat bidirectional gene organization tends to be both ancient and more conserved than alternative orientations . Similarly, through the fungi, divergent gene pairs are more conserved in orientation than convergent or co-oriented gene pairs . In some cases of divergent genes the promoter domains overlap. Here we define such bidirectional-

promoter genes as those where the nucleosome free region (NFR) of the two genes overlap. In *Saccharomyces cerevisiae* we find, in agreement with prior results, that more than 60% of non-overlapping divergent protein coding transcripts share the same promoter region. For convenience we refer to genes with a bidirectional promoter as bipromoter genes.

Bipromoter gene pairs are especially well conserved as a pair. This can be seen when comparing the current gene order in *Saccharomyces cerevisiae* with that seen in the ancestor, prior to the whole genome duplication. Comparing bipromoter pairs to divergent but non-bipromoter pairs using logistic regression, we find that bipromoter pairs are much better conserved as a pair ($p = 3 \times 10^{-7}$), even when controlling for co-

expression level ($p = 0.03$) and intergene distance ($p = 0.00136$), known predictors of pair conservation . This conservation may reflect nothing more than the fact that inversions that break up bidirectional gene pairs are more likely to disrupt promoter architecture.

Here we note that divergent orientation, bipromoter architecture in particular, is peculiar in that it puts in proximity the promoters of the two genes. This we argue may well have consequences for noise levels as for divergent genes the transcription, or priming for transcription by PolII loading, makes the transcription of the neighbour more likely, either because it might decrease the probability that the relevant chromatin stochastically closes or increases the probability of it being opened. That neighbouring genes show co-ordinated expression , that such co-ordination

is not simply owing to similarity in transcription factors and is related to local nucleosome occupancy , while noise of a transgene is dependent on the insertion site all point to a coupling between chromatin neighbourhood and noise. That transcription affects chromatin status suggests in turn that bipromoter genes are unlikely to have uncoupled expression. Indeed, in humans, it has been shown that intensive transcription at one locus frequently spills over into its physical neighbouring loci (both upstream and downstream) resulting in a time lagged burst of expression subsequent to the upregulation of the focal gene . This spill over is thought to be at least in part owing to local relaxation of chromatin associated with the expression of the focal gene, as evidenced by changes in histone modifications . The same effect is seen in yeast, only here the effect is much more highly localized, the spill over extending no further than 3kb as opposed to 100kb in humans.

Based on these observations, we propose that for bipromoter genes, the gene pair acts as

it were as a partially self re-inforcing domain of open chromatin. Such bi-promoter

domains should, we hypothesise, increase the net likelihood that chromatin is open and

should thus be conducive to low noise, enabling expression when expression is needed.

This could explain why some genes have non-coding unstable RNAs produced off a

bidirectional promoter. Below we start by examining the hypothesis by reference to

stochastic simulations.

Results:

The stochastic simulation model

Consider a pair of neighbouring genes. The promoter of each we presume can exist in

one of two states, either in open chromatin or closed. Transcription is only possible, we assume, when chromatin is open. Here, note, we ignore the possibility that transcription factors might also act to open chromatin. Assuming independent behaviour of the two genes, the probability that open chromatin closes within a fixed time interval is p_c , while the probability closed chromatin opens is p_o . If chromatin is open, then transcription is possible, occurring with a probability p_t . A transcriptional event results in N proteins before the mRNA is lost and protein decays with probability p_d .

The novel component of the simulation is to suppose that transcription of one gene might alter chromatin dynamics of the other and in turn affect transcription. There are two ways (not mutually exclusive) by which transcription of one gene might mediate

such effects: either by reducing the probability that chromatin of the other promoter will shut, if open, or by increasing the probability of the chromatin opening if shut. We model both independently and consider a third model combining both.

We start by considering the case where the probability of shutting alone is modified (model 1). We can then define a parameter, i , for the level of independence between the genes, such that if one gene is being transcribed the probability that chromatin associated with the other gene's promoter will shut will be $i.p_c$. For $i=1$, the two genes are perfectly independent (e.g. not bidirectional). For $i=0$, transcription of one gene holds open the chromatin of the other gene, if the chromatin was already open. In this model, if the chromatin of the other gene is closed, it isn't forced to open by the activity

of the neighbour. This coupling is hence in the form of resilience to chromatin closure.

In the second model, we consider that transcription of one gene increases the chances

that the promoter of the other is opened if closed, but doesn't affect the probability of

closure if open. If one gene is been transcribed, the probability that the chromatin of the

other gene will open, if closed, is $(2 - i) \cdot p_o$. In the final model (model 3), we incorporate

both effects. For further details see supplementary experimental procedures 1.

For each simulation we follow the chromatin state, the transcriptional state and the

protein level over 10,000 time units, updating status each time unit. Noise for the

protein is defined as the standard deviation in protein level over the time course / mean

level (note that variation over the time course is equivalent to variation between

unsynchronized replicates at any given time). An analogous definition is used for the

transcript-level noise. Co-expression between the two genes is the Pearson product moment correlation through the time course of the pair. Chromatin fluctuation is the probability of observing a change in chromatin state in a randomly chosen iteration.

In the model in which transcription exclusively increases the chances of closed chromatin opening (model 2), in nearly all parameter space increasing interdependence ($i \rightarrow 0$) promotes low noise. Given this, we present in detail the less permissive model (model 1). The results from models 2 and 3 are presented in Supplementary figures 1 and 2.

A typical result for model 1 is presented in figure 1. Here $p_c = p_o = 0.5$. Note that as the

likelihood of coupling decreases so noise goes up and co-expression is reduced. More generally, for a variety of parameter values we need to consider the correlation between noise level and i . If, as in figure 1, this is positive, then increased coupling, ($i > 0$), ensures reduced noise. We consider simulations in which for the two genes all parameters are the same, but we vary independently both p_c and p_o over the range 0.05 to 1 under increments of 0.05 with 10 replicates for each set of parameter values. We find that as regards transcriptional noise a positive correlation is always seen (Fig 2, blue points). However, owing to stochasticity in protein degradation this does not necessarily translate to protein level noise always decreasing with decreasing independence.

We find that protein noise can be increased when the stochastic probability of chromatin closure is high and thus most of the time no transcription is happening (Fig 2, red points). This is largely dependent on p_c not being too high (Fig 3a), as opposed to variation in p_o (Fig 3b). The causality of the negative correlation when closure probability is high is intimately related to effects on protein abundance. When closure probabilities are high (and transcription rates low), there is a positive correlation between protein abundance and protein noise (Fig 4), while, when closure rates are lower this correlation switches to a negative correlation. This most likely reflects the fact that when closure rates are high, little transcription is seen and protein levels can descend to zero, thereby reducing the variance in levels until the next transcriptional event. With some degree of coupling between the genes the protein abundance level is


raised and so noise is raised. The transcripts are, however, rare and lost almost immediately. As in yeast we see a negative correlation between protein noise and protein abundance, we surmise that true closure rates are relatively low, predicting a decrease in protein noise with increasing coupling.

Noise reduction has abundance-dependent and abundance-independent components

While in the above models we see a robust relationship between coupling and noise, much of this effect is likely to be owing to there commonly being lower noise for highly abundant proteins. In the simulations, increased coupling increases the abundance of the protein product by permitting a higher opportunity for transcription. This agrees

with the prior suggestion that, for essential genes, an increase in dose may be beneficial as it both reduces noise and moves the mean expression level away from the danger zone, where low dose equates to large fitness effects . Note too that dose sensitive genes, such as essential genes are asymmetrically dose sensitive. While reduction of dosage is very costly (hence they are deemed essential) increases in dose do not have any similar effect. Indeed, it is notable that the set of genes for which gross over-expression has a phenotype shows little overlap with the set showing fitness on reduction in dosage . We conclude that it is likely to be advantageous for some dose sensitive genes to be configured in bipromoter architecture as it increases net abundance.

Given the above logic, we might also ask whether bipromoter genes are expected to have lower noise, even allowing for the increased abundance. To approach this we consider simulations in which we alter abundance by modifying factors that affect protein abundance independent of the effects of chromatin opening and shutting, and transcriptional bursting. We can then ask whether an independent gene pair ($i=1$) and a coupled pair ($i=0$) show different noise levels when steady state protein abundance levels are equal owing to differences in decay rates (higher for coupled genes). We find for all three models that bipromoter genes still show lower noise levels at any given abundance level (Supplementary figures 3). We also consider the possibility that transcripts that result from bipromoter activity produce fewer translated proteins than do those from independent genes, keeping the decay rates constant. Again we find that

controlling for net protein abundance that coupled genes ($i=0$) have much lower noise than do independent ones (Supplementary figures 3). We conclude that noise modulation by modification of transcriptional bursting, owing to coupled gene activity, can have both abundance-dependent and abundance-independent causality. These results are in many regards comparable to those of Cook et al.  who, in examining a role for ploidy in noise modulation, identify both an abundance-dependent and abundance-independent component to noise modulation.

Bi-promoter transcribed genes have low expression noise

We tested the hypothesis that bi-promoter protein coding genes have low protein noise

with the help of recently published yeast whole genome transcription data to define gene orientation and presence of ncRNA, coupled with high resolution noise data on rich media provided for over 2000 protein coding genes specified by Newman et al. . In all, we analysed 7,272 well identified transcripts, of which 1,772 are non-coding transcripts (stable unannotated transcripts and cryptic unstable transcripts, SUTs and CUTs) which is approximately 25% of all transcripts . Among transcripts with a mapped 5' nucleosome free region (NFR), 61% of the unannotated transcripts and 48% of the protein-coding transcripts initiated bidirectionally from shared 5' NFRs rather than initiating from their own promoters .

If our hypothesis is correct, protein-coding genes with a bi-promoter architecture

(shared with either a protein coding gene or a ncRNA) should show lower expression noise. As we are not so interested in the hypothesis that bipromoter architecture might modify noise through modification of abundance, we restrict analysis to abundance corrected noise measures, as defined by Newman et al. . We also repeated analysis using residuals from a loess regression of noise against abundance and find no important differences (data not shown). After removing the confounding transcript types (5'NFR tandem transcript, 3'NFR antisense transcript and 3'NFR tandem transcript) annotated by , we find that protein-coding genes with a bi-promoter structure, sharing their 5' NFR either with a coding gene or with a non-coding gene, show significantly lower expression noise than the genes that do not have a bi-promoter transcript structure (mean noise of bi-promoter genes = 0.33 +/- 0.11; of all non-bi-

promoter genes: 1.76 ± 0.15 ; Brunner-Munzel test $p = 4.1 \times 10^{-13}$, Fig 5). More

generally, divergent genes (regardless of their NFR) have lower noise than those in

alternative configurations (noise of non-divergent genes = 1.50 ± 0.18 , mean noise of

divergent genes = 0.88 ± 0.12 , Brunner-Munzel test $p = 0.0077$). By contrast,

convergent genes don't show significant differences in noise level compared with co-

oriented genes ($p = 0.68$, Brunner-Munzel test).

Noise reduction and divergent ncRNA

This model not only has applicability in the case where both genes in the pair are

protein coding. It also has the potential to explain why some genes have antisense non-

coding RNA specified from a bi-directional domain. Such transcripts are now widely

reported. In yeast, of the unannotated transcripts (ncRNA) which have mapped 5' NFR, 61% are bidirectional initiated from a shared promoter region . Similarly, mapping millions of short RNA reads generated from murine embryonic stem cells and other differentiated cell types has revealed abundant short transcription start site-associated RNAs, many of which are antisense transcripts . Likewise in humans, depletion of the exonucleolytic RNA exosome reveals lots of highly unstable RNA of promoter upstream transcripts . Similar RNAs are reported in chicken and *Drosophila* .

One model sees these as spurious transcripts, a consequence of illegitimate transcription factor activity . Our model suggests a functional explanation. For the chromatin to remain open and for noise to be reduced, permitting expression when

expression is needed, polIII priming or transcription of a ncRNA through a promoter on

the opposite strand to that of the focal protein coding gene would be an efficient

mechanism to enable accessibility of the promoter domain of the focal gene. As

expected, we find that bi-promoter protein coding genes have low noise both when they

are partnered with a protein coding gene ($p = 5.0 \times 10^{-14}$ compared with all other genes;

Brunner-Munzel test), and when the partner is not protein coding ($p = 0.0030$, Fig. 5).

Is noise more important than co-expression?

In simulations we find that co-expression is higher when genes are coupled ($r = -0.86$).

While then the above results support the noise model, can we be confident that the

function of bi-promoter architecture is ever to reduce noise rather than to increase co-

expression levels? For the most highly co-expressed 2% gene pairs it is known that they tend to belong to the same functional class, are preserved as a pair over evolutionary time and are enriched in divergent orientation . For these there is little doubt that co-expression is functionally relevant. However, several findings support the proposition that noise modification is relevant. First, we see no significant correlation between co-expression level and mean noise, neither for divergent gene pairs ($r = -0.064$, $p = 0.424$), convergent gene pairs ($r = -0.1038$, $p = 0.152$), nor co-oriented gene pairs ($r = -0.0672$, $p = 0.257$). We do, nonetheless and as expected, find higher co-expression rates for divergent gene pairs (divergent gene pairs: mean co-expression = 0.140 ± 0.012 ; convergent gene pairs, mean co-expression: 0.107 ± 0.010 ; co-oriented gene pairs: mean co-expression: 0.101 ± 0.009 ; $p = 0.0467$ between divergent and convergent; $p =$

0.0019 between divergent and co-oriented and $p = 0.333$ between convergent and co-oriented, Brunner-Munzel test).

Second, co-presence of the product of transcription is unlikely to be the case for one class of ncRNA, cyctic unstable transcripts (CUTS), as these tend to be rapidly targeted for degradation . Importantly then, we find that when we consider protein coding genes partnered with CUTs through bi-promoters, they too have lower noise than other genes ($p=0.012$), but no different from that of protein coding genes partnered with protein coding genes in a bi-promoter architecture ($p>0.05$).

A third line of evidence derives from examination of a class of genes where *a priori* we might know the fellow genes with which they might benefit from being co-expressed.

The best candidates in this regard are proteins that belong to the same protein complex,

that do indeed have high co-expression scores with fellow members (mean co-

expression of genes from same complex: 0.1877 ± 0.0026 and mean co-expression of

genes from different complexes: 0.0253 ± 0.0001 , $p < 2.2 \times 10^{-16}$ in Wilcoxon rank sum

test). Given the need for transcription when transcription is needed, as expected

complex-associated genes do indeed have low noise ($p = 7.3 \times 10^{-7}$ Brunner-Munzel

Test). Further, as we would expect, genes specifying proteins in a complex tend to have

bipromoter architecture more than expected by chance ($p < 2.2 \times 10^{-16}$, Fisher's Exact

Test), this being true after control for essentiality ($p < 2.2 \times 10^{-16}$, Fisher's Exact Test).

While, however, complex related genes both have low noise and are found more

commonly in bipromoter architecture than expected by chance, we find no cases where

two genes specifying proteins in the same complex are located in the same bi-promoter pair. These results strongly suggest that noise modulation above co-expression is key to selection on bi-promoter genes. A very few bi-promoter genes may well also benefit from their mutual co-expression, but the more relevant force may well be selection for noise modulation.

For noise, orientation of the ncRNA matters

While above we show that ncRNA in divergent orientation is associated with low noise of the protein coding gene, this does not demonstrate that orientation *per se* is important. Is then low noise a general property of genes associated with ncRNAs, regardless of orientation, or is the divergent orientation important? We find that noise

levels of proteins with an ncRNA from the same strand as the protein coding gene have higher expression noise than proteins with a ncRNA derived from a bidirectional promoter (bi-promoter with ncRNA noise=0.65, co-oriented with ncRNA noise=2.07, $p=0.036$; Brunner-Munzel test). This both supports the hypothesis that the function of bi-promoter ncRNA is to reduce noise of the paired protein-coding gene and suggests that noise, rather than co-expression, can be the focus of selection. Moreover, genes with ncRNA from the same strand as the protein coding gene have higher expression noise than the protein coding genes which have a same strand protein coding gene neighbour ($p = 0.026$). This suggests that co-oriented ncRNAs may be a means to increase expression noise, a possibility we will not examine further.

Results are robust to covariate controls

The above results are all consistent with our hypothesis but may have alternative explanations. Previous analysis of divergent promoters in mammals suggests that several particular binding motifs are enriched in bi-promoter structures and a particular binding protein, GABP, binds to more than 80% percent of divergent promoters . This raises the possibility that differential utilization of transcription factors might explain the low noise of bi-promoter genes.

To test this, we take three transcription factors that each regulate more than 100 genes and ask whether the mean expression noise of bi-promoter genes bound by these three TFs is lower than the noise of other genes that are bound by the same TFs. Second, we

ask whether the expression noise of bi-promoter genes bound exclusively by other TFs

(i.e. not the main three) is lower than that of non-bi-promoter genes bound exclusively

by other TFs. The results show that TF binding cannot explain the low noise in bi-

promoter genes (Table 1). Further, when we control for the number of transcription

factors regulating a gene, bi-promoter genes still show lower expression noise than other

genes ($p < 0.0001$ from randomization; Supplementary Fig 5).

The existence of a TATA-box appears to be linked to increased noise levels . As bi-

directional genes in both human and *Drosophila melanogaster* often lack TATA

control, the result could reflect TATA presence/absence rather than bidirectionality *per*

se. In yeast, we find the same bias: of the 2111 protein coding genes involved in bi-

promoter pairs, only 509 are annotated as containing a TATA-box, which is

significantly lower compared to other genes ($p < 2.2e-16$, Fisher's Exact Test).

We thus compared the noise of bi-promoter TATA-containing genes with that of non-bi-

promoter TATA-containing genes, and the noise of bi-promoter TATA-less genes with

that of non-bi-promoter TATA-less genes. As expected, TATA is a predictor of noise

(e.g. in bi-promoter genes, genes with a TATA-box show higher noise levels than genes

without a TATA-box, $p = 0.0064$, Brunner-Munzel test). However, this fails to explain

the low noise of bi-promoter genes: bi-promoter genes have lower noise than non-bi-

promoter genes even when only considering those genes without a TATA-box; the same

holds when considering only genes with a TATA-box (Table 2).

Type II promoters already are nucleosome free and so don't benefit from

bidirectional architecture.

There are two types of promoter regions: those that favour nucleosomes, and those that don't . Genes with nucleosome-favoring promoters usually have high expression noise, while genes with nucleosome disfavoring promoters usually have low expression noise .

How does this relate to gene orientation?

We utilized a prior definition of type I and type II promoters . Here a type I promoter is defined as a promoter containing a TATA-box with at least 80% of the length of its binding sites covered by nucleosomes. A type II promoter is TATA-less with at most 20% of the total length of its binding sites covered by nucleosomes. We find that non-

bi-promoter genes have higher noise than bi-promoter genes when restricting our analysis to nucleosome-favouring promoters (>80% occupancy; mean noise =1.81 in bi-promoter genes, noise=5.46 in other genes, $p = 0.00020$, Brunner-Munzel test; Table 3).

This remains true after controlling for gene essentiality (Table 3).

By contrast, for genes with nucleosome-disfavouring promoters (occupancy <20%), we see no evidence for a noise reduction through bi-promoter architecture (Table 3). If

Seila et al are correct this result is to be expected. They conjecture that RNAPII

complexes are simultaneously engaged at the boundaries of the nucleosome-depleted

region surrounding TSSs and that these divergently engaged polymerases could directly

reinforce the -1 and +1 nucleosome positions, effectively enhancing the boundaries of

the nucleosome-free region, allowing transcription factors access to the promoter , and maybe further maintaining the “loose” chromatin during transcription . Such genes are, in effect, primed for transcription, regardless of orientation: an ‘interrupted form’ of bi-directional transcription occurs even if there is no bi-promoter. For those bi-promoter pairs that do not exclude nucleosomes in this manner from the bi-promoter region during transcription (type I pairs), dependence between the two genes is re-inforced and noise reduced, much as we modelled. If the above picture is true, we would expect that the class II (nucleosome-free) genes in non-bidirectional orientation should have lower noise than class I genes in the same orientation, which indeed we observe (mean noise level is 0.10 +/- 0.17 and 5.46 +/- 0.70, respectively. $p < 2.2e-16$, Brunner-Munzel test; this remains true when controlling for gene essentiality). In short, nucleosome depletion

and bidirectional orientation we suggest to be two alternative mechanisms to ensure low noise by resisting stochastic chromatin closure.

Only bi-promoter genes show correlated noise of neighbours

For any gene we can assay its noise level under a variety of parameter values. The simulation suggests that when two genes are coupled ($i > 0$) the noise levels of the two proteins across these multiple conditions are correlated. More generally, across all simulations we consider the correlation in protein noise between the neighbours for a given value of independence i . We find this to be strongest when coupling is strongest ($r = -0.96$). Our simulations thus predict that the correlation in noise levels between neighbours should be strongest when coupling is strongest and hence when genes are

divergent. If independence of divergent genes is in turn modulated by intergene

distance, by the same logic we expect for divergent genes the correlation in noise levels

to be higher when intergene distance is lower.

Confirming these predictions, we find a significant correlation of the noise of two

divergent transcripts. Conversely, neither convergent nor co-oriented gene pairs show

correlated noise levels (Spearman rank correlation for divergent pairs $r = 0.148$, $p =$

0.031 ($r = 0.151$, $p = 0.047$ after removing type II genes); for convergent pairs $r =$

0.0089 , $p = 0.45$; for co-oriented pairs $r = -0.0008$, $p = 0.51$; p -values determined by

randomization). Also as predicted the mean noise level of the transcripts in divergent

gene pairs is correlated with the distance between transcription start sites, a correlation

not seen for convergent and co-oriented pairs (Spearman rank correlation for divergent pairs $r = 0.0936$, $p = 0.0055$; for convergent pairs $r = -0.0194$, $p = 0.49$; for co-oriented pairs $r = -0.0282$, $p = 0.29$).

Essential genes tend to be low noise with bi-promoter architecture, while the opposite is seen for stress response genes.

Of all genes, those that are lethal on knockout (i.e., essential) are most likely to be under selection for reduced noise levels . Conversely, stress related genes are thought to be under selection for high noise . Many features of essential genes are consistent with low noise. They tend to be highly expressed, but even controlling for this they have low noise . Counter-intuitively for highly expressed genes the mRNAs have short half lives ,

a feature consistent with low noise . They tend not to be TATA controlled and reside

clustered in genomic low noise/open chromatin domains .

If bi-promoter architecture is a mechanism to enable low noise and expression when

needed, we might also expect such genes to be in divergent or bi-promoter orientation

more than expected by chance. This is indeed the case in yeast. Of 6600 protein coding

genes in yeast, 2627 are divergent with a partner protein coding gene. Of these, 537

(20.4%) are essential, while only 577 (14.5%) of the 3973 non-divergent genes are

essential. There is thus enrichment of essential genes in the divergent class ($p = 4.9 \times$

10^{-10} , Fisher's exact test). There is a corresponding enrichment of essential genes in

gene pairs with bi-promoter architecture. Of 2111 genes in bi-promoter organization,

22% are essential, while only 649 of 4489 (14.4%) non-bi promoter genes are essential

($p=5.9 \times 10^{-13}$, Fisher's exact test). An analogous excess in divergent orientation has

recently been reported in *Drosophila*. Moreover, we see more bidirectional pairs of two

essential genes than expected by chance: there are 79 bidirectional essential gene pairs

in yeast, this being more than ever found in 1000 gene order randomizations, $p<0.001$).

Also as expected, haploinsufficient genes tend to be in bipromoter architecture more

than expected (41% versus 31% of all others; $p=0.005$).

For stress-related genes, where we expect selection for high noise, we see the opposite

pattern. While those that are bi-promoter have lower noise than stress related genes in

different configurations (mean noise for bi-promoter stress genes 1.59 ± 0.30 , for non-

bi-promoter stress genes 3.63 ± 0.27 , $p=1.6 \times 10^{-8}$, Brunner-Munzel test), stress related genes tend to avoid having a bi-promoter architecture. Only 509 (24.1%) bi-promoter genes are stress related, while 1525 (34.0%) of non-bi-promoter genes are stress related (Fisher's exact test, $p = 2.7 \times 10^{-16}$). Similarly, stress genes tend not to be in divergent orientation (28% divergent, 32.5% non-divergent; $p = 0.00024$, Fisher's exact test).

What of the essential genes that are not bi-promoter with another protein coding gene?

We predict to see more cases of antisense ncRNA than expected by chance associated with such genes, if ncRNA is a mechanism of noise reduction. This we observe. Of 309 genes with an antisense CUT, 65 (21%) are essential genes, while only 624 (14.1%) of 4441 genes without an antisense CUT are essential ($p = 0.0014$, Fisher's exact test).

If there are peculiar features of essential genes (e.g. short half life, low usage of optimal codons), can we exclude the possibility that bi-promoter genes have low noise just because of this enrichment for essential genes? Mean noise level of the 1646 non-essential bi-promoter genes is significant lower than other non-essential genes (0.43±0.12 versus 1.83±0.15, $p = 5.5 \times 10^{-12}$ in Brunner-Munzel test). That non-essential genes with bi-promoter control have lower expression noise than essential genes (in all orientations) ($p = 0.035$) further suggests that dispensability cannot alone account for the low noise of bi-promoter genes.

There must, however, be alternative methods to modulate noise. Notably, we find that the mean noise of bi-promoter essential genes (with either an ncRNA or a protein

coding gene partner) is not significantly lower than the noise of non-bi-promoter

essential genes (0.18 ± 0.22 versus 0.22 ± 0.26 , $p=0.82$ in Brunner-Munzel test;

0.03 ± 0.20 versus 0.29 ± 0.30 , $p = 0.76$ after removing type II genes). These results are

then consistent with bi-promoter architecture being a means to reduce noise, but,

unsurprisingly, not the only mechanism.

What the other mechanisms might be is not immediately transparent. For example,

while essential genes have a shorter mRNA half life than non-essential genes ($p = 2.8 \times$

10^{-16} , Brunner-Munzel test), the mean mRNA half life for bi-promoter essential genes is

no different to that of non-bi-promoter essential gene (16.65 versus 16.91 respectively: p

$= 0.25$, Brunner-Munzel test). Increased usage of codons that specify abundant tRNAs

is expected to enable fast translation and be associated with high noise. As expected, there is a positive correlation between the frequency of optimal codon usage (FOP) and expression noise in yeast ($r = 0.107$, $p = 4.6 \times 10^{-07}$, Spearman's rank correlation).

However, FOP of bi-promoter essential genes does not differ from that of either essential non-bi-promoter genes or essential non-divergent genes ($p = 0.16$ and 0.63 , respectively, Brunner-Munzel tests).

Bi-promoter gene pairs and CUTs are rare in noisy subtelomeric domains

Does the fact that bi-promoter gene pairs have low noise affect not only which sort of genes are found in this architecture but also where on chromosomes they are found?

Previously it was reported that essential genes and non-essential genes flanked by a

high density of essential genes tend to have low noise . Could it be that non-bi-promoter essential genes tend to reside in essential gene clusters, thus giving them low noise? Alternatively might genes requiring low noise not only adopt bi-promoter architecture but also aggregate into low noise chromosomal domains? Ignoring genes +1 and -1 from a focal essential gene (direct neighbours) and then asking about the number of essential genes in the flanking 5 genes on either side, we find that both bi-promoter essential genes ($p=0.022$) and bi-promoter non-essential genes ($p=0.018$) have more essential genes in their vicinity than expected by chance (Table 4). Thus bi-promoter genes tend to be enriched in the vicinity of essential gene clusters, these having unusually low noise levels . Clustering of bipromoter genes doesn't however fully account for the low noise of genes in such domains. Examining non-bipromoter

genes, those in essential gene clusters have lower noise than those not in clusters

($P=0.0007$; controlling for essentiality, $P=0.01$).

Yeast subtelomeric domains are high-noise domains and are depauperate in essential

genes. From the logic that bi-promoter architecture is a genomic device to minimize

noise, we might expect that genes found in subtelomeric domains should be favoured to

be high noise genes and hence not in a bi-promoter architecture. Considering all genes,

28 of 324 gene pairs (8.6%) are bi-promoter in subtelomeric domains (20kb from

chromosome ends), while 2083 of 6276 (33%) non-subtelomerics are bi-promoter

($p < 2.2 \times 10^{-16}$, Fisher's exact test). However, as essential genes tend to be bi-promoter

and avoid subtelomeric domains, we may be seeing nothing more than the biased

distribution of essential genes. Considering only non-essential genes, we see the same bias (8% subtelomeric non-essential genes in bi-promoter architecture versus 31% non-subtelomeric, $p < 2.2 \times 10^{-16}$, Fisher's exact test). We similarly find that bi-promoter CUT associated genes are rare subtelomerically (1.2% subtelomeric genes have a bi-promoter CUT compared with 4.8% otherwise, $p=0.001$ Fisher's exact test; this remains when controlling for essentiality of the neighbour, $p=0.006$). The high noise of subtelomeric genes and the avoidance of subtelomeric domains by bipromoter genes cannot explain the low noise of bipromoter genes, as they have low noise even compared with genes that are not subtelomeric ($P < 10^{-11}$).

Discussion

We have found, via simulation, that if transcription of one gene increases the probability of transcription of a neighbour and vice versa, then low noise of both is expected across broad and realistic parameter space. We propose that divergent gene pairs, bi-promoter gene pairs in particular, are thus expected to be low noise genes, even allowing for any effect on protein abundance. This model has striking predictive ability. Bi-promoter genes are indeed low noise and, as predicted, the noise is modulated by intergene distance. Similarly, bipromoter pairs have correlated noise. The model can predict biases both in which genes are or are not in bipromoter architecture (essential/complex genes and stress response genes respectively) and which classes of gene should be more likely to have ncRNA in bi-promoter architecture. Indeed, that our model can predict noise levels and skew in gene type associated with CUTs, strengthens the view that

noise control, independent of co-expression modulation, is a focus of selection. The model also predicts that bipromoter pairs should be rare subtelomerically as observed, such domains being high noise domains.

These results suggest that gene orientation may well be an important feature in the control of noise, they also suggest that, as with transcription at SER3 , it is the act of transcription, rather than the product of transcription, that can be important. While the CUT associated with SER3 (a sense transcript) is associated with control of the expression of the downstream gene, we argue that transcription from the opposing strand is an effective mechanism for priming a focal sense strand gene for expression and hence for reduction in noise. The transcript may well be unwanted, but it doesn't

follow that the making of the transcript is without functional relevance. This is also supported by the observation that upstream RNA PolIII transcripts usually cannot be elongated effectively .

We might then also wonder how much expression in protein coding genes from bidirectional promoters is to enable noise control rather than produce the protein product itself. Such a hypothesis could explain why many relatively highly co-expressed neighbours ($0.4 > r > 0.2$) in yeast have no functional (GO class) similarity .

These findings add to recent evidence that a substantial component of selection on gene arrangement within genomes is to modulate noise levels. In yeast the clustering of essential genes may be owing to such selection (see also). In bacteria co-linearity, the

tendency for genes to appear in the same order in the operon as the proteins are needed in a temporal fashion, appears also best explained by the consequences of selection on noise . What remains to be resolved is whether noise modulation mediated by changes in gene order/orientation is relevant in less compact genomes, such as those of mammals.

Materials and methods:

Dataset

All yeast (*Saccharomyces cerevisiae*) transcripts as observed by tiling arrays under three conditions (YPE, YPD and YPGal) and their genomic coordinates were obtained from . Two transcripts were considered as bi-promoter transcripts if they share the same 5' nucleosome free region (NFR), where NFR was defined as a nucleosome deplete

region ≥ 80 bp, according to . These transcripts were defined as divergent (), convergent () or co-oriented (or) by their coordinates in the genome. Essential genes in rich media were downloaded from the web site of the *Saccharomyces* Genome Deletion Project (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html). Both the yeast gene order (Version 2) and genome annotation information were taken from (<http://wolfe.gen.tcd.ie/ygob/>). For more than 2,000 proteins, expression noise data in rich media were obtained from . We used the distance to median noise level (DM_YEPD) in our analysis to get rid of the confounding influence of protein abundance. Genes whose promoter contains a TATA-box were derived from a large TATA-box gene enquiry experiment . Codon usage bias (FOP) was obtained from . The

relationships between transcription factors (TF) and their target genes were derived from the yeast transcriptional regulatory network . In total, 12873 regulatory interactions were identified in this network. Stress-related genes and growth-related genes were obtained from and co-expression level of adjacent gene pairs as previously reported . Haploinsufficient genes were taken from and Genes with type I and type II promoters were obtained from . 431 type I genes and 565 type II genes were included in our analysis. Protein complexes were gained from .

Data analysis

Transcripts that share the same 5' NFR were described in Xu et al. . The noise of each protein measured by Newman et al. was used to represent the noise of the transcript. In

the comparison of the noise of proteins derived from divergent transcripts to the noise of proteins without divergent transcripts, transcripts with complex annotations were excluded (e.g. the annotation “other”, which means the transcript contains multiple open reading frames or is a mixture of non-coding and coding parts). In the calculation of the correlation between noise levels of protein pairs, transcripts that contain multiple annotation features (e.g. the annotation “other”, which means the transcript contains multiple open reading frames or is a mixture of non-coding and coding parts) were excluded. In the calculation of the correlation between noise level and the distance between transcription start sites, we used the mean noise level of the two proteins if the noise of both proteins had been measured. If one gene transcript shares its promoter with a non-coding transcript, the noise of this gene was chosen to represent the noise of

the two transcripts in the calculation. We used the lawstat package in R to perform the

Brunner-Munzel test .

Randomization test of the correlation between noise levels of gene pairs.

Our model predicts that the expression noises of two divergent genes should be

positively correlated due to the shared chromatin regulation, as chromatin regulation

processes are responsible for much of the expression noise in yeast . To check if there is

a positive correlation between expression noise in divergent, convergent and co-oriented

gene pairs, and to obtain the significance level of any such correlation, we employed a

randomization procedure. In this we extract the noise level for each protein, orient the

gene pairs by their strand location for divergent and convergent gene pairs, by their

transcription order for co-oriented gene pairs, calculate the spearman correlation level

for this data, randomize one column of genes 10,000 times and determine the

correlation for each. The significance level of the observed correlation is $(m+1)/10001$

where m is the rank of the true correlation compared against the randomizations.

Randomization test to determine whether essential-essential gene pairs are more likely to be divergent gene pairs.

The *S. cerevisiae* gene order was taken from the Yeast Gene Order Browser

(<http://wolfe.gen.tcd.ie/ygob/>), Version 2. The procedure is as follows: 1: count the

number of divergent essential gene pairs in the *S.cerevisiae* genome; 2. randomize the

position of essential genes in each chromosome 1,000 times and calculate the number of

divergent essential gene pairs for each; 3. The significance level of this number is

$(m+1)/1001$, where m is the rank of the true number compared with the randomizations.

Method to test to the density of essential genes in different gene types.

To calculate the density of essential genes surrounding essential bi-promoter genes and

essential non-bi-promoter genes, a +/- 5 gene window was used to scan the yeast

chromosomes (the *S. cerevisiae* gene order we used is from

<http://wolfe.gen.tcd.ie/ygob/>, as described above). To avoid biases caused by the fact

that essential genes tend to be in divergent gene pairs, the direct (+1 and -1) gene

neighbors were excluded from the scan.

References

Acknowledgments

We thank Claudia C. Weber, Lu Chen for helpful discussion and Araxi O. Urrutia for sharing her yeast co-expression data.

Author contributions. LDH and GZW conceived the work, performed the analysis and wrote-up the work. All authors contributed to writing up and proposed tests.

Funding. GZW is supported by a Boehringer Ingelheim travel grant and by the visiting research scholar program of the University of Bath. LDH is a Royal Society Wolfson Research Merit Award Holder. The funders had no role in study design, data collection

and analysis, decision to publish, or preparation of the manuscript.

Competing interests. The authors have declared that no competing interests exist.

Figure Legends

Figure 1. The relationship between the independence between two neighbouring genes and various noise and co-expression parameters. For this plot $p_c=p_o=0.5$.

Data: transcriptional noise, blue; protein noise, red; co-expression, green; chromatin

fluctuation, black; proportion of time chromatin open, grey. Other parameter values:

$N=100$, $p_r=0.9$, $d=0.7$.

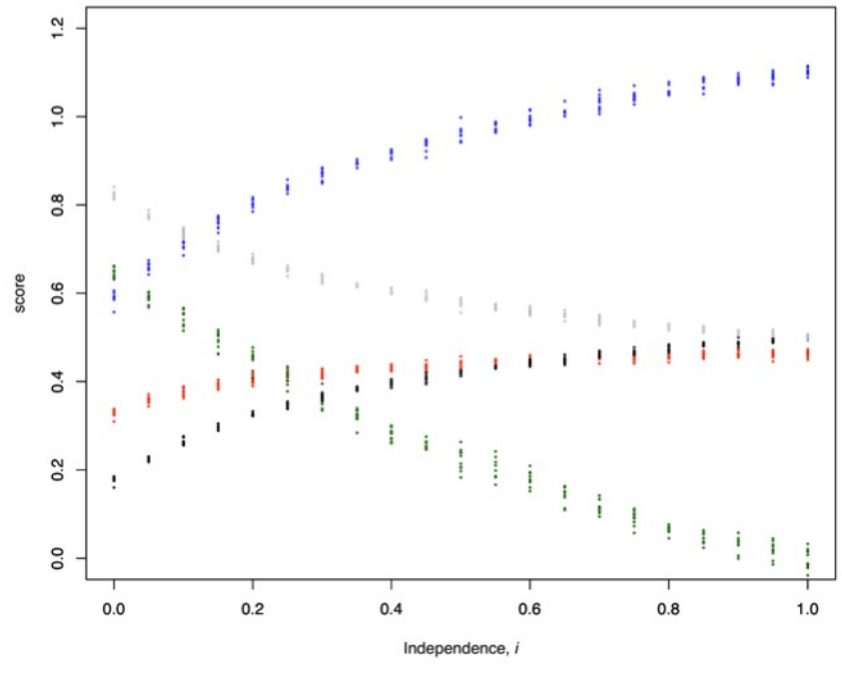


Figure 2. The correlation between noise and independence as a function of the ratio of the probabilities of chromatin opening and shutting. A positive correlation

indicates decreased noise with increasing inter-dependence. Protein noise, red;

transcriptional noise, blue. Other parameter values, $N=100$, $p_f=0.9$, $d=0.7$.

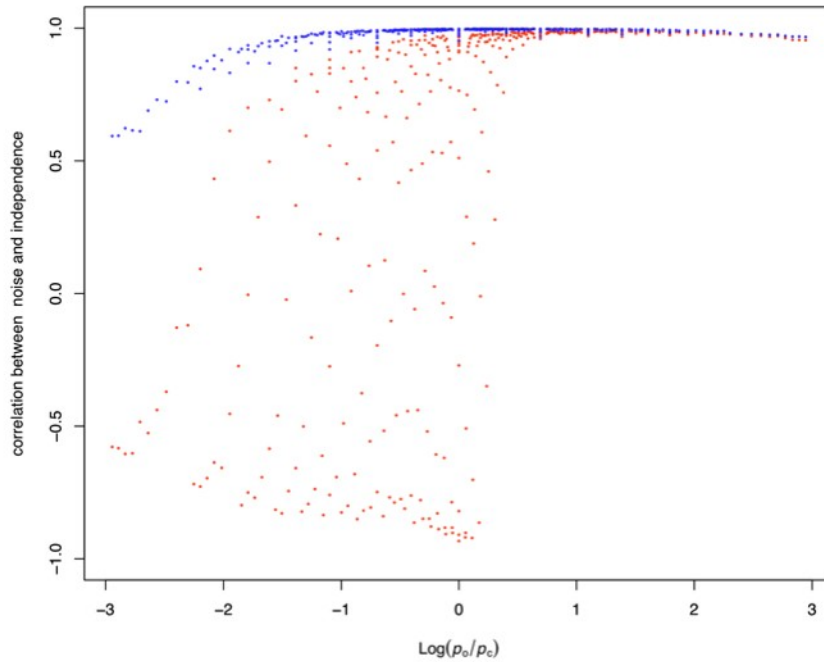
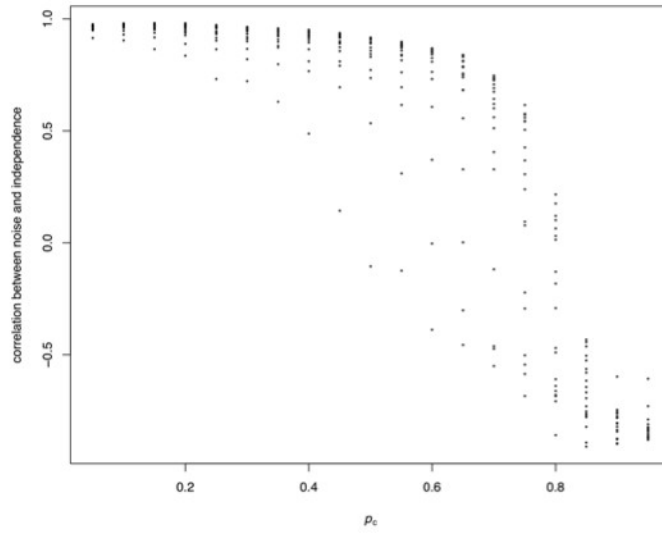


Figure 3. The correlation between noise and independence as a function of the probabilities of chromatin shutting (3a) and opening (3b). A positive correlation indicates decreased noise with increasing inter-dependence. Protein noise, red; transcriptional noise, blue. Other parameter values, $N=100$, $p_f=0.9$, $d=0.7$.

a.



b.

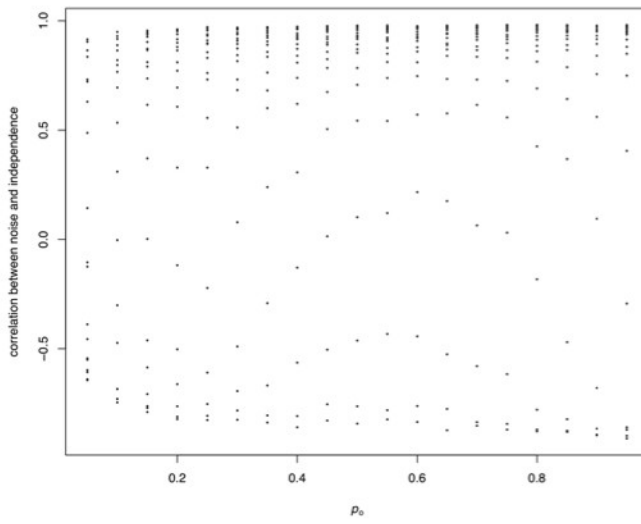


Figure 4. The relationship between the correlation between protein noise and protein abundance as a function of the probability of chromatin closure.

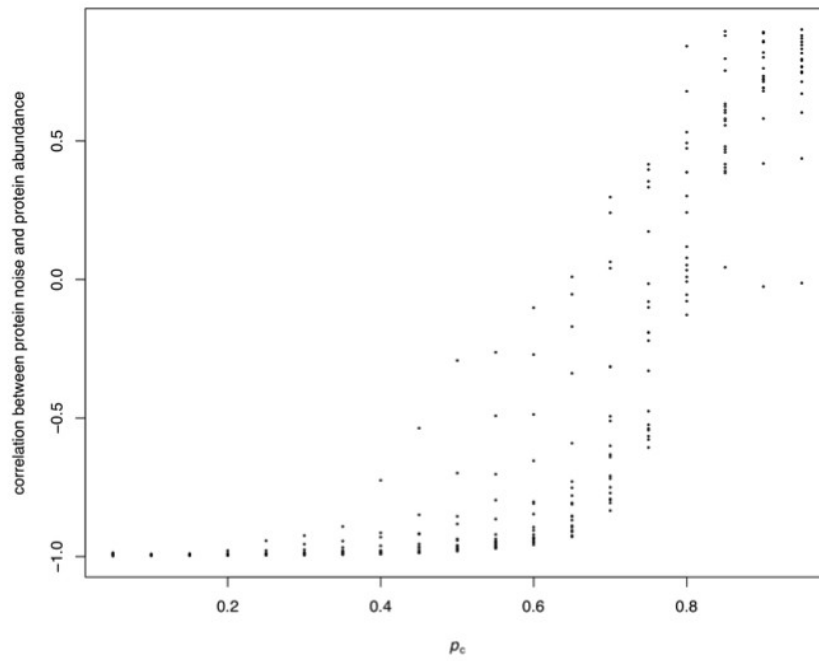


Figure 5. Genes which share a promoter (5' NFR) with either a non-coding transcript or coding transcript (ORF) show lower expression noise than genes without any bi-promoter transcript. Number of genes that have noise value in each categories, With non-coding: 216; With orf: 537; Other (genes that do not share 5'NFR with other transcript): 1072. In this plot, the boxes are drawn with widths proportional to the square-roots of the number of observations in the groups". Non-overlapping notches on the boxes are roughly equivalent to non-over sem error bars.

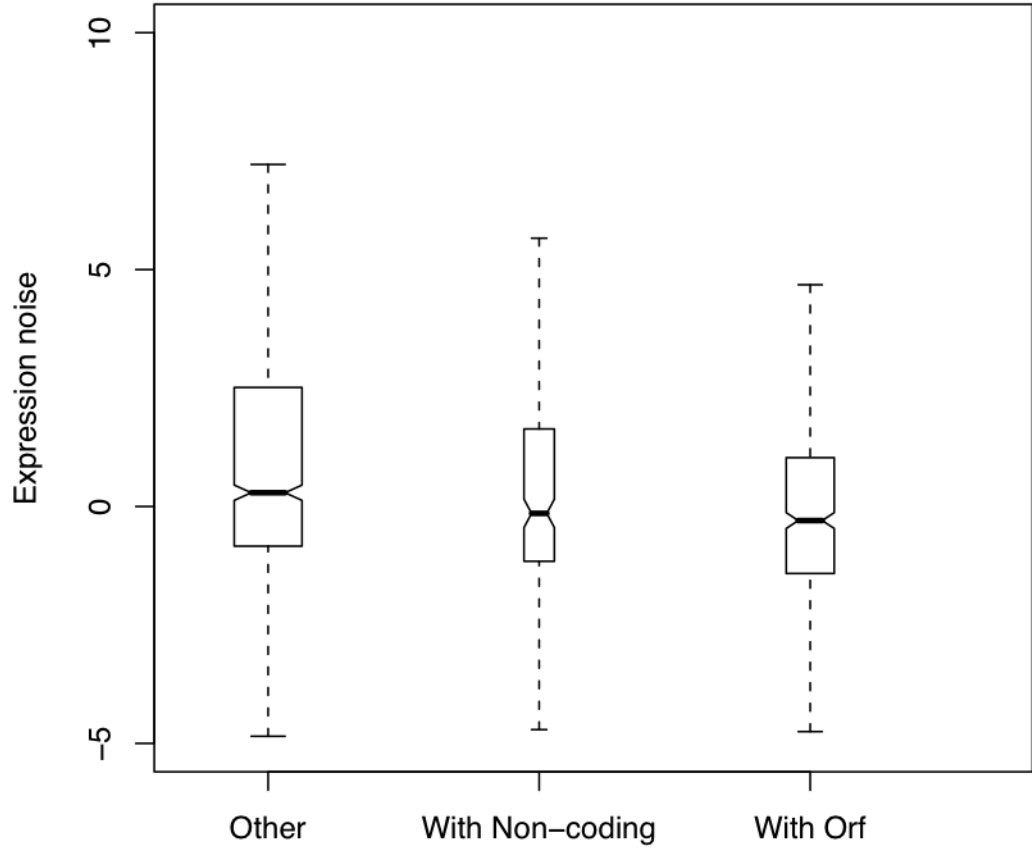


Table 1. Binding of particular transcription factors cannot explain the low noise of

bi-promoter genes. The noise level of bi-promoter genes is significantly lower than that

of other genes both in the case of genes regulated by the same common transcription

factor, and for those regulated by other transcription factors. *p*- values from Brunner-

Munzel tests.

	Regulated by particular TFs	Regulated by other TFs
Bi-promoter genes	0.09 +/- 0.24 (322)	0.43 +/- 0.12 (1789)
Non-bi-promoter genes	1.15 +/- 0.26 (568)	1.57 +/- 0.15 (3921)
p-value	0.0013	4.0e-08

Table 2. The low noise of bi-promoter genes cannot be explained by TATA boxes.

Noise levels of bi-promoter genes are significantly lower than those of other genes, both

in genes with TATA box containing promoters in TATA-less genes. Mean noise+/-

standard error (number of genes). *p*- values from Brunner-Munzel tests.

	TATA box-containing genes	TATA-less genes
Bi-promoter genes	1.01+/-0.25 (509)	0.13+/-0.11 (1602)
Non-bi-promoter genes	2.71+/-0.27 (1587)	0.70+/-0.12 (2902)
<i>p</i> -value	2.1e-06	0.0013

Table 3. Nucleosome favouring bi-promoter genes have lower noise than

nucleosome favouring non-bi-promoter genes. In the control for essentiality we just

examine the non-essentials. *p* value determined by the Brunner-Munzel test.

	nucleosome favouring	nucleosome disfavouring
bi-promoter	1.81 +/- 0.66 (103)	0.09 +/- 0.30 (233)
non bi-promoter	5.463 +/- 0.698 (328)	0.10 +/- 0.17 (331)
<i>p</i>	0.00020	0.16
control for essentiality	nucleosome favoured	nucleosome disfavoured
bi-promoter	2.491 +/- 0.829 (87)	-0.2801 +/- 0.2018 (162)
non bi-promoter	6.1434 +/- 0.756 (301)	0.188 +/- 0.203 (241)
<i>p</i>	0.0028	0.065

Table 4. The density of essential genes among the 10 genes flanking focal genes. Here

we ignore genes +1 and -1 of a focal gene (direct neighbours).

	Bi-promoter	Not bi-promoter	<i>p</i>-value
Essential	0.212 +/- 0.006	0.195 +/- 0.005	0.022
Not essential	0.188 +/- 0.003	0.180 +/- 0.003	0.018
<i>p</i> -value	0.00089	0.010	