Title: Transcription factor binding activity is the primary driver of noise in gene expression

Authors

Lavisha Parab[1†], Sampriti Pal[1†], Riddhiman Dhar[1]*

Affiliations

[1] Department of Biotechnology, Indian Institute of Technology (IIT) Kharagpur, Kharagpur, West Bengal-721302, India.

*email: riddhiman.dhar@iitkgp.ac.in

†These authors contributed equally to this work

Abstract

Noise in expression of individual genes gives rise to variations in activity of cellular pathways and generates phenotypic heterogeneity. Phenotypic heterogeneity has important implications for antibiotic persistence, mutation penetrance, cancer growth and therapy resistance. Even though the role of specific factors such as presence of TATA box sequence in the promoter of a gene and promoter nucleosome occupancy in noise regulation have been studied, the ability of these factors to predict noise remains unknown. Here we build an integrated quantitative statistical model for gene expression noise in yeast and show that the number of regulating transcription factors and their binding activity are the primary drivers of noise. Combining our model with stochastic simulations, we specifically show that the binding activity arising from competition and cooperation among TFs for binding to the promoter region can drive noise. Our work uncovers a general process of noise regulation that arises out of the dynamic nature of gene regulation that is not dependent on specific transcription factor or specific promoter sequence and thus can be extended across biological systems.

Running title: A dynamic origin of expression noise

Keywords

Gene expression noise, Cellular processes, Integrated model, Transcription factors, Cooperative and Competitive binding

Introduction

Random fluctuations in molecular events operating inside a cell generate variations in the expression levels of genes which leads to variation in activities of cellular pathways. Such variations ultimately impact cellular phenotypes and generate phenotypic heterogeneity among individual cells of an isogenic population under identical environmental condition. Phenotypic heterogeneity has many important biological implications. Heterogeneity in cell growth rate and gene expression has been shown to give rise to persister cells that are able to survive antibiotic treatment (Balaban *et al.*, 2004; Rotem *et al*., 2010; Wakamoto *et al*., 2013; Arnoldini *et al*, 2014; Page and Peti, 2016). Phenotypic heterogeneity can also facilitate emergence of antibiotic resistant populations (Levin and Rozen, 2006; Bergmiller *et al.*, 2017; Levin-Reisman *et al.*, 2017; Dewachter *et al.*, 2019). Stochastic variation in expression level of genes can give rise to variable fitness effect of a mutation among individuals of a population – a phenomenon known as incomplete penetrance (Eldar *et al.*, 2009; Raj *et al*., 2010; Burga *et al.*, 2011; Dickinson *et al.*, 2016; Taeubner *et al.*, 2018), which has important implications in development of multi-cellular organisms and in human diseases (Dickinson *et al.*, 2016; Taeubner *et al.*, 2018). In addition, phenotypic heterogeneity has a key role in growth and evolution of cancers (Meacham and Morrison, 2013; Nguyen *et al.*, 2019; Sharma *et al.*, 2019) and has also been associated with emergence of drug resistance in cancer (Pisco *et al*, 2013; Sheng *et al.*, 2018; Hammerlindl and Schaider, 2018; Gupta *et al.*, 2019; Farquhar *et al.*, 2019).

Stochastic variations in the expression of genes have been systematically quantified in microbial systems such as bacteria *Escherichia coli* and yeast *Saccharomyces cerevisiae* (Newman *et al.*, 2006; Bar-Even, A. *et al.*, 2006; Taniguchi *et al.*, 2010; Silander *et al.*, 2013) and the molecular origins of noise have been investigated in many studies over the years (McAdams and Arkin, 1997; Elowitz *et al.*, 2002; Blake *et al.*, 2003; Raser and O'Shea, 2004; das Neves *et al.*, 2010; Sanchez *et al.*, 2011; Hornung *et al.*, 2012; Salari *et al.*, 2012; Sanchez *et al.*, 2013; Sharon *et al.*, 2014; Chen and Zhang, 2016; Wu *et al.*, 2017; Faure *et al.*, 2017; Baudrimont *et al*., 2019). These studies have shown a correlation between the presence of TATA box motif in the promoter sequence of a gene with higher expression noise (Bar-Even *et al.*, 2006; Tirosh *et al.*, 2006; Hornung *et al.*, 2012; Ravarani *et al.*, 2016). Further, promoter nucleosome occupancy, alone as well

as in conjunction with presence/absence of TATA box motif, has also been associated with expression noise (Tirosh and Barkai, 2008; Choi and Kim, 2009; Small *et al.*, 2014). Recent studies have also associated histone modification patterns with expression noise (Weinberger *et al.*, 2012; Chen and Zhang, 2016; Faure *et al.*, 2017; Nicolas *et al.*, 2018). These factors can influence transcriptional burst size and frequency (Donovan *et al*., 2019; Larsson *et al*., 2019; Nicolas *et al*., 2018) which, in turn, impact gene expression noise (Hornung *et al.*, 2012; Zoller *et al.*, 2015; Wang *et al*., 2019; Engl *et al*., 2020). However, even after so many years, how well these molecular factors can predict noise remains completely unknown and a predictive model of noise remains elusive. Such a model can greatly boost our capabilities to decipher noise in genes across biological systems. In addition, this can also help in designing synthetic promoters with desired noise levels for driving gene expression in an organism.

Further, even though noise in expression in individual genes has been well-studied, the architecture of noise in cellular pathways has remained largely unexplored beyond the traditional gene-set enrichment analysis. A majority of the cellular traits and phenotypes are polygenic in nature, meaning they are shaped by coordinated functioning of groups of genes. Fluctuations in the expression of such group of genes shape the variations in the activity of cellular pathways, thereby generating phenotypic heterogeneity. Thus, it is essential to decipher the pattern of noise at the gene and pathway levels to quantitatively understand variations in cellular phenotypes.

In the current work, we build an integrated statistical model of gene expression noise in yeast by combining all possible molecular factors that can impact noise. We test the predictive ability of each of these molecular factors and uncover that the transcription factor binding activity has the highest predictive power. We show that the number of regulatory transcription factors (TFs) of genes in a cellular process impacts noise. Further, we show that an increase in number of regulatory TFs results in overlaps in TF binding sites and leads to competition between TFs for binding to the same genomic region. Through mathematical modeling and stochastic simulations, we demonstrate that either cooperation or competition between TFs for binding

to the promoter can increase noise. Further, we systematically quantify noise in biological processes and show that transcription factor binding activity is also the best predictor of noise in biological processes. Our work through a combination of an integrated, quantitative and predictive statistical model of noise and mathematical modeling uncovers a general mechanism of noise regulation that is not dependent on the specific details of the biological system. This mechanism instead highlights a dynamic origin of noise and can thus be extended across biological organisms.

Results

We considered gene expression noise both at the level of mRNA and protein. For calculating gene expression noise at the level of mRNA, we used the single-cell RNA-seq data in yeast from Nadal-Ribelles et al. (2019). This data showed expected scaling of noise (coefficient of variation, CV) with mean mRNA expression level (Fig. 1A), where noise increased as the mean mRNA expression level decreased. We calculated a measure of noise that was independent of mean expression level through fitting a polynomial curve to the noise vs mean plot and calculating vertical distance of CV values from the fitted line. We obtained the expression noise at the protein level from the data generated by Newman et al. (2006) (Fig. 1A) and used their measure of distance to median (DM) as the measure for expression noise, which was again independent of mean protein level. Noise values obtained at the mRNA level was well correlated with the noise at the protein level ($r=0.39$, $p=1.7\times10^{-61}$, Fig. 1B).

To quantify the power of molecular factors to predict noise, both at the mRNA and protein levels, we separated the noise data into training and test datasets (Fig. 1C). The training dataset consisted of 80% of the whole data and the test set contained the remaining 20% data. To fit linear regression models on the noise data, we first identified the key factors after removing multicollinear variables through Ridge or Lasso regression. We then fitted a regression model on the training data and analyzed the fraction or percentage of variation that the model could explain. We then used the fitted model to make predictions on the test data and computed predicted R^2 value. Higher values of fraction variation explained and predicted R^2 values

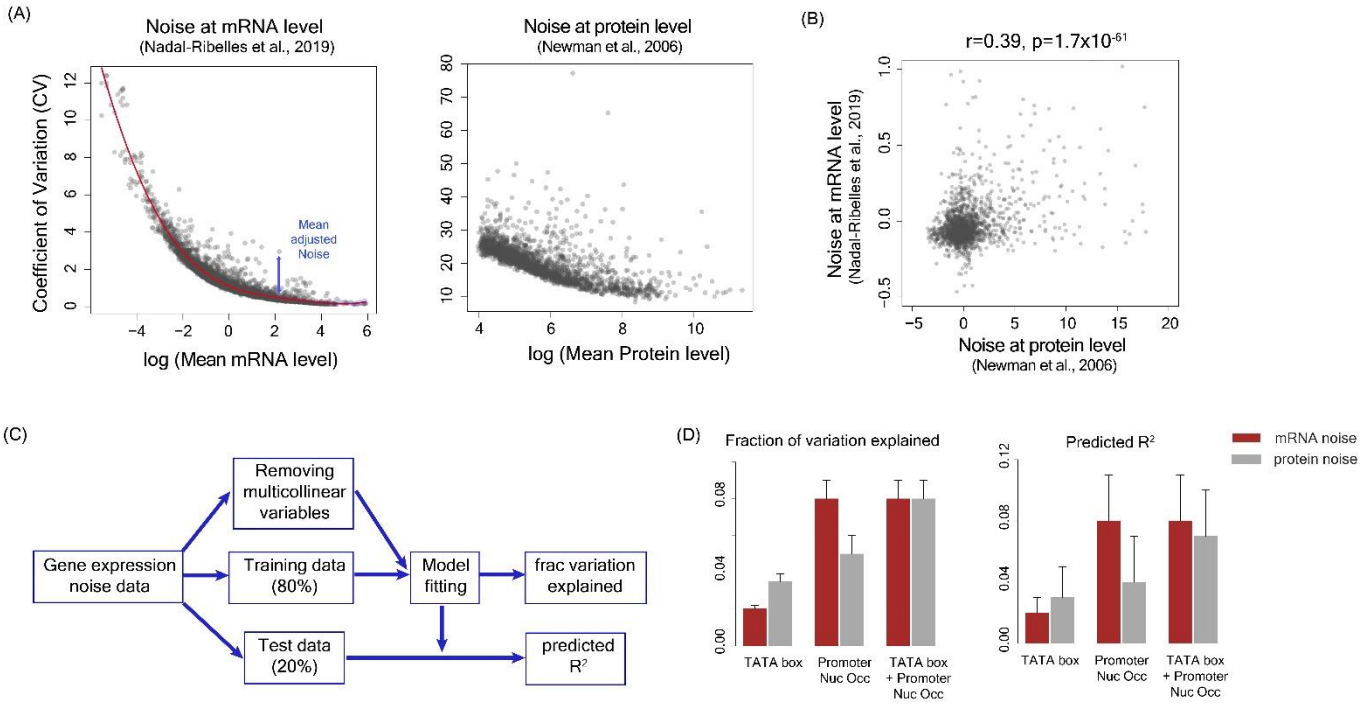
meant that a model is a better predictor of noise than the models with lower values for these measures. We performed this analysis in mRNA noise dataset as well as in protein noise dataset to ensure that our findings were not biased by a specific dataset.

Molecular factors that have been associated with expression noise, such as presence of TATA box sequence in the promoter (Bar-Even *et al.*, 2006; Tirosh *et al.*, 2006; Hornung *et al.*, 2012; Ravarani *et al.*, 2016) and promoter nucleosome occupancy (Tirosh and Barkai, 2008; Choi and Kim, 2009; Small *et al.*, 2014) were poor predictors of noise (Fig. 1D), even though they showed some correlation with noise (Supplementary fig. S1). Specifically, TATA box sequence, promoter nucleosome occupancy alone and in conjunction with TATA box sequence, could only explain ~2-4%, ~5-8% and ~8% of the noise variation in biological processes respectively and had low predictive power (predicted R^2 value 0.02-0.03 for TATA box alone, 04-0.08 for promoter nucleosome occupancy alone, and 0.07-0.08 for TATA box +promoter nucleosome occupancy; Fig. 1D). This suggested that these factors were largely associative with noise and not predictive.

To systematically identify the key molecular factors that could predict noise, we built an integrated statistical model considering all factors that are known or are likely to influence gene expression (Fig. 2). Transcription factors are key regulators of gene expression and thus can have significant impact on noise. We considered the number of transcription factors binding to the promoter region of a gene, location of their binding sites, their mean expression and noise levels in our model. In this regard, we also considered the SAGA or TFIID dependence of genes for their expression as SAGA dependent genes has been shown to be activated by stress whereas TFIID dependent genes are largely housekeeping genes (Huisinga and Pugh, 2004). We also considered whether a gene is co-activator redundant or TFIID dependent from the classification by Donczew *et al.*, 2020. We further considered the binding activity of several broadly acting

TFs such as TBP, ABF1 and RAP1 (van Werven *et al.*, 2009; Lickwar *et al.*, 2012; de Jonge *et al.*, 2020). In addition, histone modifications strongly influence gene expression, thus we considered histone levels and histone modification patterns in the promoter region, in the transcription factor binding sites as well as in the gene body in the integrated model. Further, we considered three-dimensional genomic contacts within a chromosome and among chromosomes, as such contacts can also influence gene expression. We also included the binding patterns of chromatin remodelers in our model (Yen *et al.*, 2012; Zentner and Henikoff, 2013; Ramachandran *et al.*, 2015).

Figure 1 - TATAbox sequence and promoter nucleosome occupancy are poor predictors of gene expression noise. (A) Left – Noise values calculated at the mRNA level from single cell RNA-seq data in



yeast (Nadal-Ribelles *et al.*, 2019). The mean adjusted noise was calculated by fitting a polynomial curve to the CV vs mean plot and calculating vertical distance of each noise value from the fitted curve. Each point show CV and mean mRNA level values for a gene. **Right -** Noise values calculated at the protein level from flow cytometry measurements by Newman *et al.*, 2006. **(B)** Noise at the mRNA level was well correlated with noise at the protein level (r=0.39, p=1.7x10^{-61}) **(C)** Flowchart showing the steps for model fitting, calculation of fraction of variation explained and predicted R^2. **(D)** Fraction of variation in noise explained and predicted R^2 value by presence or absence of TATA box sequence, average promoter nucleosome occupancy per nucleosome bound site and the combination of presence or absence of TATA box sequence with promoter nucleosome occupancy.

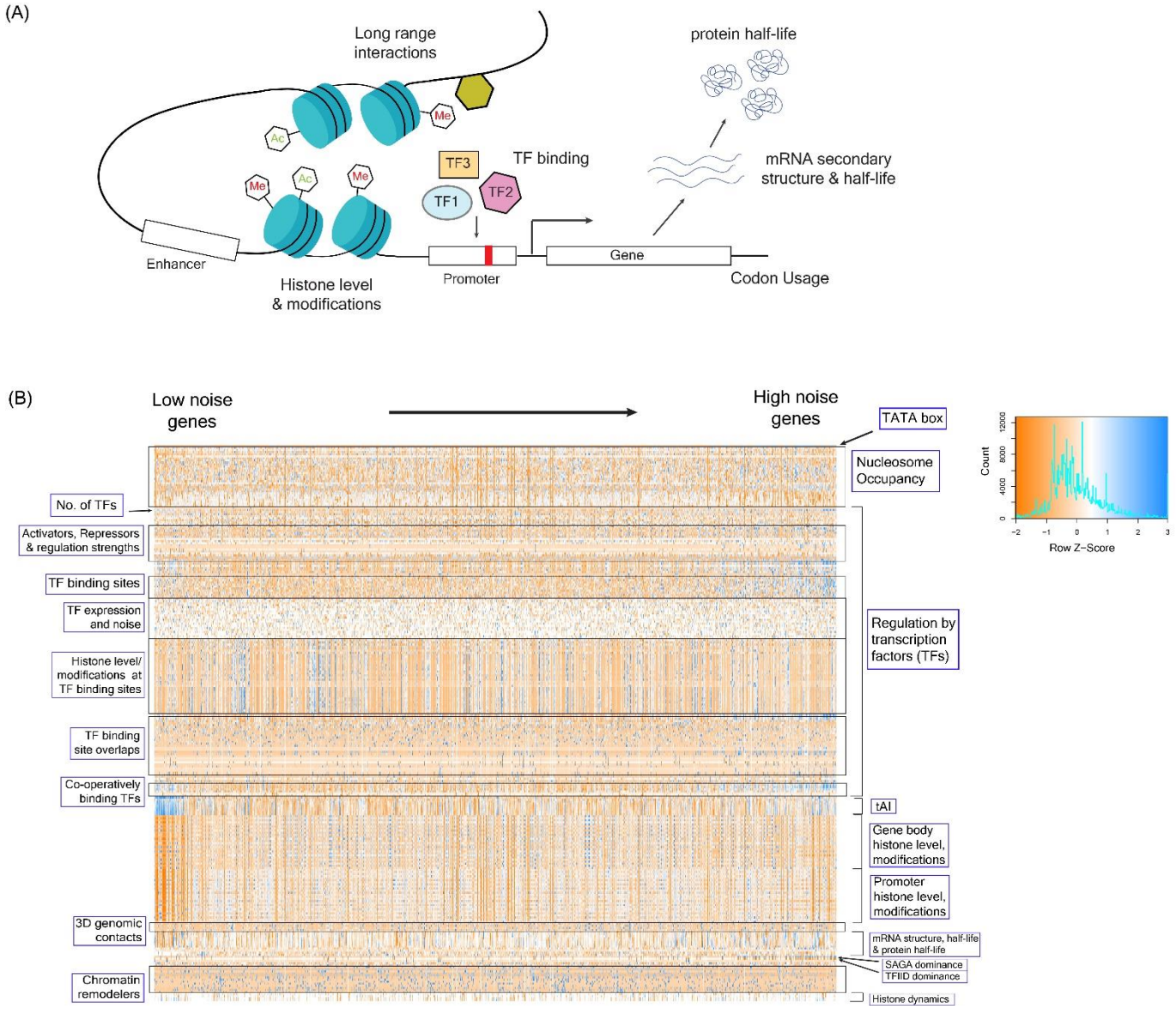


Figure 2 – An integrated statistical model of gene expression noise (A) Schematic diagram depicting the possible molecular factors that could impact gene expression and thus, could have a key role in expression noise. **(B)** An integrated model of noise in biological processes constructed considering TATA box sequence, absolute nucleosome occupancy levels, gene regulation by transcription factors, tRNA adaptation index, histone modification pattern in gene-body and promoter region, 3D genomic contacts, mRNA structure and half-life, protein half-life, activity of chromatin remodelers and histone binding dynamics. The heatmap shows values of all these factors (scaled and centered) in genes (represented in the columns) sorted according to noise level.

Furthermore, we also took tRNA adaptation index into account as this can impact the translation process and can impact noise (Salari *et al.*, 2012). Finally, we also considered mRNA secondary structure, mRNA half-life and protein half-life as these factors can influence the protein level in a cell (Kertesz *et al.*, 2010;

Geisberg *et al.*, 2014; Belle *et al.*, 2006). We considered all these factors in addition to nucleosome occupancy pattern and presence and absence of TATA box sequence in the promoter region that have been considered before. For nucleosome occupancy, we not only considered the number of nucleosome-bound sites in the promoter region and gene body but also included the absolute nucleosome occupancy pattern obtained from Oberbeckmann *et al.* (2019) (Fig. 1D) and the histone binding dynamics obtained from (Dion *et al.*, 2007) (Fig. 2).

We tested each factor for its ability to explain variation in the noise data and to predict noise, for both mRNA and protein noise datasets. We then ranked these factors according to their ability to explain the variation present in these datasets (characterized by fraction of variation explained) and by their ability to predict noise (characterized by predicted R^2 value). The ranking of the factors, whether based on fraction of variation explained or based on predicted R2 value were highly correlated among mRNA and protein noise datasets (Fig. 3A, B).

The top 10 factors for explaining the variation in noise data and for predicting noise contained the same factors even though their ranking were slightly different. Interestingly, eight of these factors were associated with the transcription factor binding, suggesting a key role for transcription factors in noise regulation (Fig. 3C). These included number of regulatory TFs for a gene (fraction of variation explained ~10% and a predicted R^2 of ~0.1) and the number of transcription factor binding sites (fraction of variation explained ~11-12%, predicted R^2 ~0.11-0.12). The remaining two factors were related to SAGA-dependence and TFIID-dependence of genes for their transcription. SAGA dependence and TFIID dependence could explain 10-19% and 9-17% variation respectively with predicted R^2 values of 0.1-0.18 and 0.09-0.17 respectively (Fig. 3C). However, these values differed greatly between mRNA and protein noise datasets (Fig. 3C).

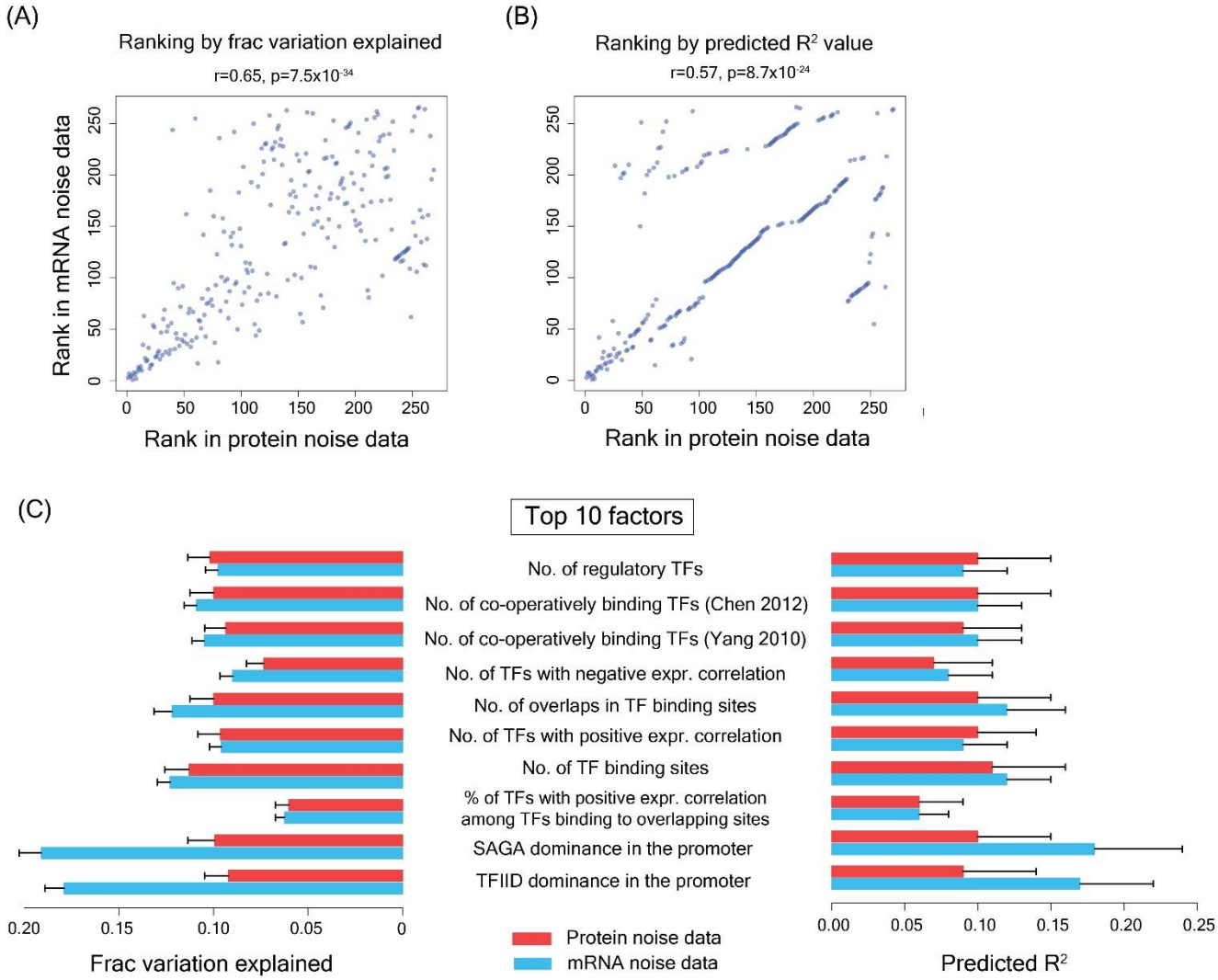


Figure 3 - Factors that can explain the variation in noise and have high predictive power are largely related to transcription factor binding activity (A) Rankings of factors according to the fraction of variation explained in mRNA noise dataset and in protein noise dataset are highly correlated (r=0.65, $p=7.5\times10^{-34}$) **(B)** Rankings of factors according to the predicted R^2 value in mRNA and protein noise datasets are highly correlated (r=0.57, $p=8.7\times10^{-24}$) **(C)** Fraction of variation explained and predicted R^2 value for top 10 factors for both mRNA and protein noise datasets.

To validate our findings, we further classified genes into high and low noise categories according to their noise values. At the mRNA level, a gene was considered to be highly noisy if the mean adjusted noise value was greater than 0.3 and a gene was assigned low-noise category if the mean adjusted noise value was less than -0.3 (Supplementary fig. S2A). This classified 386 genes as high-noise and 157 genes as low-noise genes. At the protein level, a gene was assigned high-noise category if the DM value was greater than 1 and

a gene was assigned low-noise category if the DM value was less than -1 (Supplementary fig. S2B), classifying 498 genes as high-noise and 387 genes as low-noise genes. Interestingly, considering noise at the mRNA level, ~66% of the high noise genes were SAGA dependent where none of the low noise genes were SAGA dependent (Fig. 4A). Similarly, for noise at the protein level, ~32% of the high-noise genes were SAGA dependent where only ~5% of the low-noise genes were SAGA dependent. Conversely, for noise at the mRNA data, all low-noise genes were TFIID dependent whereas only ~31% of the high-noise genes were TFIID dependent. Similarly, for noise at the protein data, ~93% of the low-noise genes were TFIID dependent whereas ~66% of the high-noise genes were TFIID dependent (Fig. 4A).

Genes with high expression noise are regulated by higher number of transcription factors

Surprisingly, our model revealed a significant correlation between number of regulating transcription factors and noise both at the mRNA (Supplementary fig. S3A) and protein level ($r=0.52$, $p=6.2\times10^{-15}$; Fig. 4B). The genes in high-noise processes on average had >30% more regulatory transcription factors compared to the genes in the low-noise processes ($p=8.9\times10^{-15}$ for mRNA noise data and $p=1.3\times10^{-15}$ for protein noise data, Fig. 4B). This increase in number of regulatory transcription factors resulted in an increase in number of TF binding sites upstream of the start codon (Supplementary fig. S4A). In addition, genes in high-noise processes showed an upstream shift in the point of lowest nucleosome occupancy (Supplementary fig. S4B), suggesting variation in TF binding patterns as a likely driver of noise.

The increase in number of regulatory TFs led to an increase in the number of TFs that bound co-operatively to the promoter region (Yang *et al*., 2010; Chen *et al*., 2012), with >60% increase in the number of TFs showing co-operative binding in high-noise genes compared to that in low-noise processes ($p<10^{-16}$ for both mRNA and protein noise data; Fig. 4C; Supplementary fig. S5A). In addition, genes with high noise showed an increase in the number of co-expressing TFs that were positively correlated in their expression ($p<10^{-14}$

both for mRNA and protein noise; Fig. 4D). Further, high-noise genes also had higher number of TFs that showed negative expression correlation among themselves (Supplementary fig. S5B). These results suggested possible roles for co-operatively binding TFs and co-expressing TFs in noise regulation.

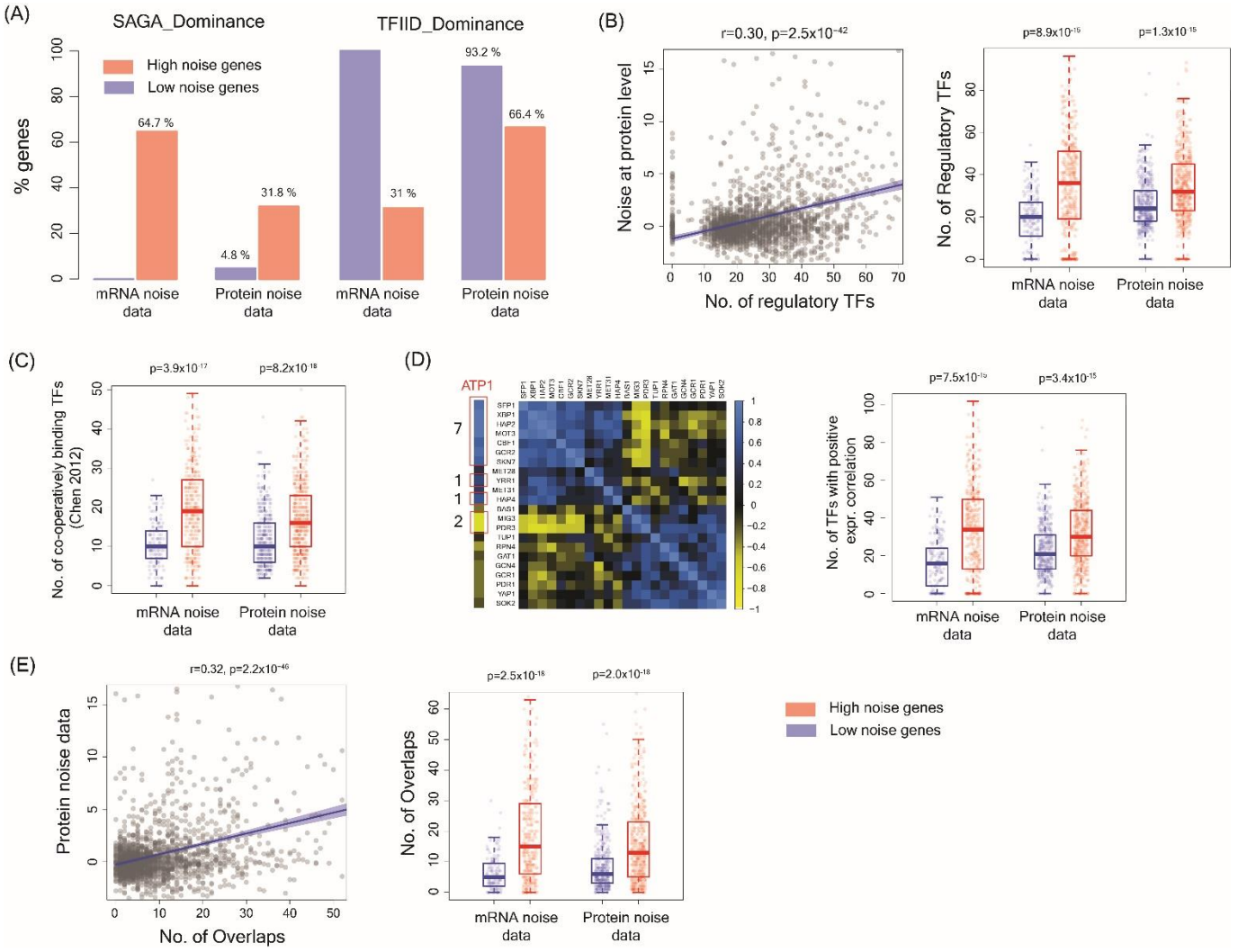


Figure 4 - Genes with high noise are regulated by higher number of TFs, have higher number of cooperatively binding TFs, and have more overlaps in TF binding sites compared to low-noise genes. (A) Percentage of genes showing SAGA and TFIID dominance in high- and low-noise genes across mRNA and protein noise datasets **(B) Left -** Correlation between noise at the protein level and number of regulatory TFs of genes. **Right** – Difference in the number of regulatory TFs between genes with low and high noise, for noise calculated at the mRNA level as well as protein level **(C)** The percentage of co-operatively binding TFs for low- and high-noise genes. **(D) Left** - Expression correlation between a gene (ATP1) and all its regulatory TFs identifies activating and repressing TFs of the gene. The significant positive and negative correlations for the gene ATP1 are highlighted in red at the left of the matrix. The correlation matrix depicts the expression correlation within TFs of the ATP1 gene. **Right** - the number of TFs that show significant positive expression correlation (co-expressing) among themselves for low- and high-noise genes. **(E) Left** - Correlation of noise at the protein level with the number of overlaps in TF binding sites. **Right** - Number of overlaps per TF binding site for low and high-noise genes.

Overlaps in TF binding sites lead to higher noise

Interestingly, an increase in the number of regulatory TFs and a corresponding increase in their binding sites led to a substantial increase in overlap of TF binding sites in the promoter region. The number of overlaps could explain ~10-12% of the variation in noise data and could predict noise with predicted R^2 value of 0.1-0.12. The median number of overlaps increased by over 100% in high-noise genes compared to low-noise ones ($p<10^{-17}$ for both mRNA and protein data; Fig. 4E). In high-noise genes, most of the increase in the overlaps of TF binding sites were observed at distal sites beyond 500 bp upstream region of the start codon (Supplementary fig. S5C). We further looked into the number of TFs that were expressing at the same time and also bound to overlapping binding sites in the promoter region, as these were the TFs that were likely to compete for binding to the same site in the promoter region. We did this by calculating the number of overlaps in binding sites of each pair of TFs as a fraction of the total number of TFs for a gene. Interestingly, genes in high noise processes showed >50% increase in percentage of co-expressing TFs whose binding sites had overlap ($p<10^{-9}$ for n=both mRNA and protein noise data; Fig. 4F) compared to genes in low-noise processes.

Overlaps in binding sites of co-expressing TFs causes competition between these TF molecules for binding to the same region of the promoter. TF binding is a dynamic process consisting of rapid binding and unbinding steps (Burger *et al*., 2010; Das *et al*., 2017), and binding competition can generate noise in two possible ways. First, competition between TFs leads to a scenario where a gene is regulated by different subsets of TFs at different points of time. This generates temporal variation in TF binding (Fig. 5A). In presence of TFs that differ in their strengths of regulation, such temporal variation in TF binding leads to temporal variation in transcription rate in each individual cell. Second, asynchronous temporal variation in TF binding among individual cells in a population leads to inter-individual variation in TF binding (Fig. 5A).

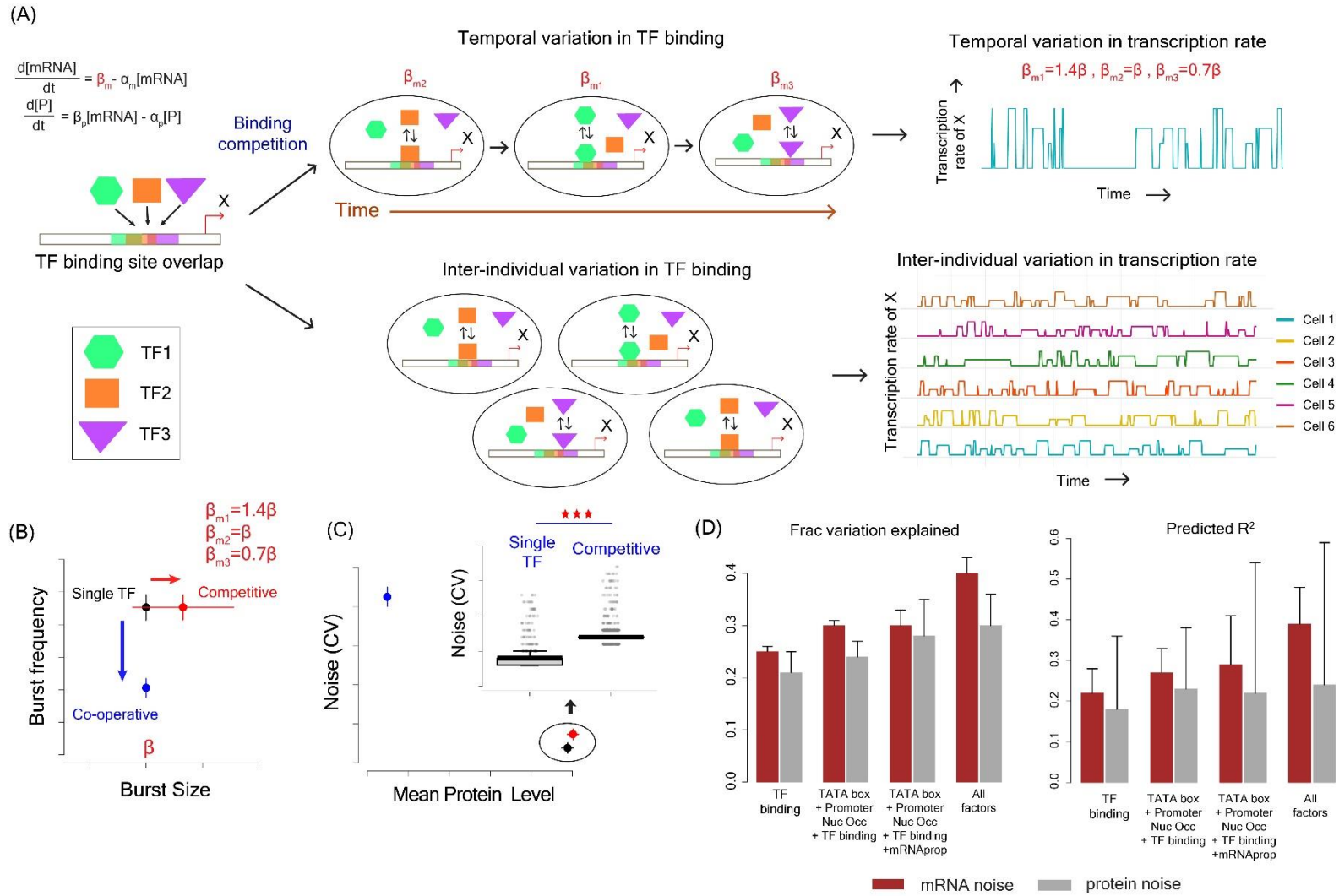


Figure 5 - Binding competition for overlapping TF binding sites drive noise. **(A)** Overlap between TF binding sites lead to binding competition between TFs. This gives rise to temporal variation in TF binding in the same promoter region within a cell. In addition, asynchrony in competitive TF binding among individual cells gives rise to inter-individual variation in TF binding and transcription rate. **(B)** Mathematical modeling and simulations show that co-operative TF binding changes the burst frequency but may not change the burst size, whereas competitive TF binding with TFs of varied regulation strengths alters burst size but may not affect burst frequency. **(C)** Simulations show that both co-operative and competitive TF binding changes noise (CV). Inset - noise distribution in cases of gene regulation by single TF binding and competitive TF binding calculated across multiple time-points in 10,000 simulated cells. **(D)** Fraction of variation explained in gene expression noise data and predictive ability (given by predicted R2 value) by factors associated with TF binding activity, combination of TF binding activity with other factors and by the combination of all factors.

We compared the impact of co-operative and competitive binding of TFs to promoter region on noise through mathematical modeling and stochastic simulations (Supplementary fig. S6A). Co-operative binding required co-ordination between two or more TFs which led to a reduction in burst frequency even though burst size remained the same (Fig. 5B, Supplementary fig. S6B). In contrast, competitive TF binding did not alter burst frequency but changed average burst size as well as introduced higher variability in burst size when regulated by TFs of different strengths (Fig. 5B). At the protein level both co-operative and competitive binding of TFs led to higher noise (Fig. 5C, Supplementary fig. S6B).

Overlaps in binding sites for different TFs can also lead to degeneracy of TF binding sequences to accommodate different consensus binding motifs of different TFs. Such degeneracy can change binding affinity of TFs to the DNA and can lead to noisy transcription. However, we did not see any difference in binding site degeneracy among genes of high and low- noise genes (Supplementary fig. S7).

As expected, all aspects of TF binding activity taken together was the single most important contributor and made up the largest part of the faction of variation explained and predictive power of all the factors combined. The integrated model comprising of all factors was able to explain ~40% of the variation in noise at the mRNA level and ~30% of the variation in noise at the protein level (Fig. 5D). TF binding activity alone explained 25% of the variation in the noise at the mRNA level and 21% of the noise at the protein level (Fig. 5D). In addition, the integrated model was able to predict the noise at the mRNA level with predicted R^2 value of 0.39 and the noise at the protein level with predicted R^2 value of 0.24 (Fig. 5D). As before, TF binding activity alone could predict noise at the mRNA level with predicted R^2 value of 0.22 and the noise at the protein level with predicted R^2 value of 0.18 (Fig. 5D).

Regulatory TF activity also predicts noise in biological processes

Noise in biological processes - shaped by the noise levels of individual genes - has important implications for phenotypic heterogeneity. Thus, we investigated whether the TF activity could predict noise in biological processes. To quantify noise in biological processes, we analyzed the overall noise levels in biological processes based on Gene Ontology (GO) classification in yeast (Ashburner *et al.*, 2000; Carbon *et al.*, 2019). We calculated median noise for each process by considering expression noise of all genes involved in the process (Fig. 6A). Noise value for each individual gene was obtained from mean adjusted noise value calculated from single-cell RNA-seq data (Nadal-Ribelles *et al*., 2019) as well as from distance to median (DM) measure calculated for protein noise data (Newman *et al., 2*006). In addition, we estimated noise dispersion in each process by calculating interquartile range of the noise values of all genes in the process. We considered only those processes for which we obtained noise values for 10 genes. This allowed us to calculate noise values for 475 biological processes based on mRNA noise data and for 192 processes based on protein noise data. The median noise and noise dispersion values obtained from these datasets were well correlated (Supplementary fig. S8).

Biological processes showed a wide range of distributions for median noise values and noise dispersion (Fig. 6B and Supplementary fig. S9). Interestingly, majority of the biological processes (>50%), however, did not have any significant deviation towards high or low noise and were balanced in their content of high- and low-noise genes (Wilcoxon rank-sum test, FDR<0.1). Processes showing high noise both based on mRNA noise data and protein noise data included cellular pathway associated with oxidative stress response. Common processes that exhibited low noise in both mRNA and protein noise datasets were associated with transcription and translation.

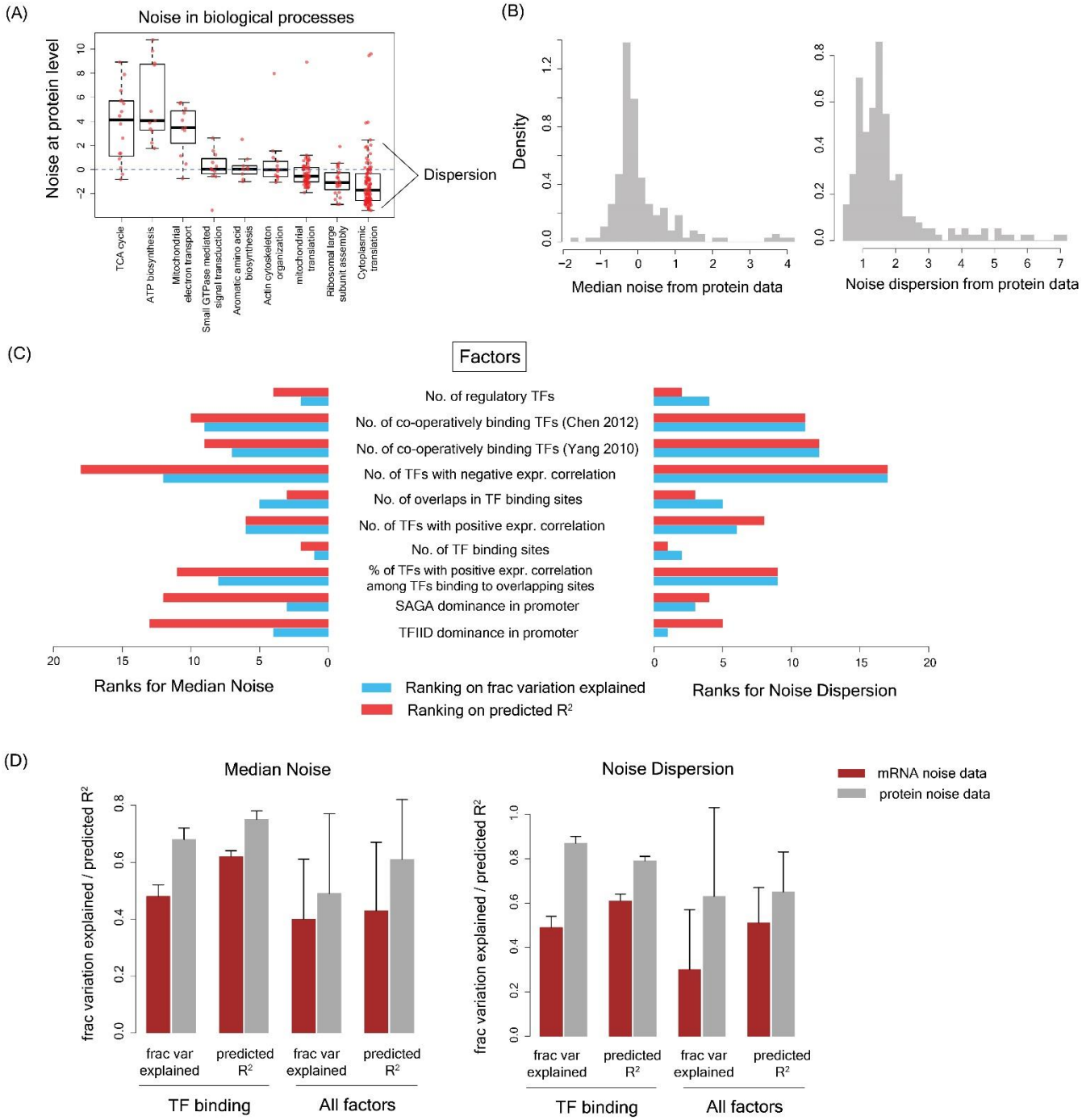


Figure 6 - TF binding activity also predicts noise across biological processes. (A) Examples of noise distribution in several biological processes with noise calculated at the protein level. **(B)** Distribution of median noise and noise dispersion across biological processes in the yeast *Saccharomyces cerevisiae*, calculated from the noise at the protein level **(C)** Ranking of factors based on the fraction of variation explained and by predictive ability given by the predicted R^2 value for median noise and noise dispersion across biological processes. **(D)** Fraction of variation explained in noise across biological processes and predictive ability (given by predicted R^2 value) by factors associated with TF binding activity, and by the combination of all factors.

Next, we checked how each of the factors used in the analysis of noise at individual gene level could explain the variation and could predict noise across biological processes. We tested the ability of each of these factors to explain and predict process noise calculated from both mRNA and protein noise data and ranked them. Interestingly, all of the top 10 factors for explaining and predicting noise at the gene level could also explain and predict noise of the biological processes to a large extent (Fig. 6C). These factors could also explain and predict noise dispersion across biological processes (Fig. 6C). In addition, most interestingly, the number of regulatory TFs and the number of TF binding sites were among the top predictors for both median noise and noise dispersion across biological processes (Fig. 6C).

Again, TF binding activity was the single most important contributor to the faction of variation explained and predictive power of all the factors combined. The integrated model comprising of all factors was able to explain 62-75% of the variation in median noise value depending on whether noise was calculated based on mRNA or protein noise (Fig. 6D). TF binding activity alone explained 48-68% of the variation in the median noise (Fig. 6D). In addition, the integrated model could predict median noise with predicted R^2 value of 0.43-0.61 whereas the TF binding activity alone could predict median noise with predicted R^2 value of 0.4-0.49. Further, the integrated model was able to explain 61-79% the noise dispersion of the biological processes and was able to predict noise dispersion with predicted R^2 value of 0.51-0.65 (Fig. 6D). As before, TF binding activity alone could explain 49-87% of the variation in noise dispersion across biological processes and was able to predict noise dispersion with predicted R^2 value of 0.0.3-0.63 (Fig. 6D). These results highlighted a pivotal role of the regulatory TF activity in predicting noise across biological processes.

Discussion

In summary, through an integrated quantitative analysis we have shown that the gene expression is predominantly explained and predicted by the transcription factor binding activity. Noisy genes tend to be controlled by a larger number of TFs. These include a substantial fraction of TFs that bind co-operatively to the promoter region and the coordination required between these TFs for binding can drive noise. In addition, increase in number of TFs leads to overlapping TF binding sites which can lead to competition between TFs for binding to the same promoter region. This can give rise to a temporal variation in TF binding within a cell and can generate noise when TFs of different strengths of regulation are present. In addition, such competition can generate inter-individual differences in TF binding pattern and in presence of TFs of varied regulation strengths can generate noise. Although competition between specific interacting partners of TATA binding protein has been implicated in noise before (Ravarani *et al.*, 2016), we uncover a much-more general molecular mechanism of noise generation that is not dependent on any specific TF or any specific promoter sequence.

Although our integrated model can predict a substantial fraction of noise variation, there is, however, still a large fraction of noise that cannot be explained by our model. This could be because of two reasons. Firstly, it is possible that several other molecular factors which can generate noise have not been accounted for in our model. Some of these molecular factors may have not been discovered yet. Second, the inherent randomness of molecular processes might not allow us to completely predict the noise variation. This aspect of noise could thus limit our predictive capability.

Finally, our findings provide a step forward for prediction of noise in genes and cellular processes across biological organisms. Recent explosion in genomics data has led to genome-wide characterization of TF binding sites across a diverse range of organisms. Thus, the knowledge of TF binding sites maps and

overlaps can enable us to predict noise in the regulated genes and subsequently in the associated cellular processes. In addition, with increasing availability of genome-wide nucleosome occupancy maps, histone modification patterns and three-dimensional genome configuration data, our study also provides a framework for building integrated model of gene expression and its noise across a diverse range of organisms. Stochastic variations in molecular processes are ubiquitous in cells across biological systems and have major implications in human diseases and their treatment including cancer evolution and drug resistance emergence. Thus, increased ability to predict variations in biological processes will be extremely useful in predicting the extent of heterogeneity in cellular traits and phenotypes.

Methods

Noise calculation in the expression of individual genes

Noise values for individual genes in yeast at the protein level was obtained from Newman *et al.*, 2006 and the DM values in the SD medium were used for all noise analysis. The DM values in the SD medium were highly correlated with DM values measure din the YPD medium and did not affect our results (data not shown). The noise values of all genes at the mRNA level were calculated from the single-cell RNA-seq data provided by Nadal-Ribelles *et al.*, 2019 as follows. Briefly, for each gene the coefficient of variation (CV) was calculated from its mean expression and the standard deviation value. Different polynomial fits were made to the CV vs log-transformed mean expression value and the best fit was chosen. A polynomial of order 5 was found to give the best fit. The mean adjusted noise value for a gene was obtained by calculating the vertical distance between the CV value and the best fitted line.

Noise analysis in biological processes

Biological processes were classified according to the Gene ontology classification for *S. cerevisiae* obtained from http://geneontology.org/ (data version: releases/2017-11-22) (Ashburner *et al.*, 2000; Carbon *et al.*,

2019). Gene names were mapped to GO terms using SGD yeast gene association map (https://www.yeastgenome.org/, GOC Validation Date: 11/10/2017) (Cherry *et al.*, 2012). Only terms associated with Biological Processes (BP) were included in our analysis and if gene expression noise data for at least 10 genes were available for that class. For each included GO term, median noise value and noise dispersion (through calculation of inter-quartile range) were calculated. To identify processes with high or low noise, one-sample Wilcoxon signed-rank test was performed to check whether the noise values of genes in a biological process differed significantly from zero followed by multiple hypothesis testing correction using Benjamini-Hochberg method (Benjamini and Hochberg, 1995). A False Discovery Rate (FDR) cut-off of 10% was used for identifying high and low-noise processes.

Building an integrated model of noise in yeast

The integrated model of noise was generated by carefully considering all factors that can regulate gene expression. These included several aspects of TF binding dynamics, presence of TATA box sequence in the promoter, absolute nucleosome occupancy levels in the promoter and the gene body, histone modification patterns in the promoter and the gene body, variations in transcription start sites (TSS), tRNA adaptation index (tAI) and three-dimensional contacts in the genome (Fig. 1D). Promoters with the presence of TATA box sequence were identified from the data of Rhee and Pugh, 2012. For each biological process, percentage of genes with TATA box sequence in their promoters was calculated. The genome-wide absolute nucleosome occupancy level for yeast genome was obtained from Oberbeckmann *et al*., 2019 and average nucleosome occupancy level per nucleosome occupied site for all promoter regions and gene bodies were calculated. The region from 1000bp upstream to 10bp downstream of the start codon was considered to be the promoter region of a gene. Average nucleosome occupancy level per occupied site calculated between -1000bp to -900bp region of the start codon was shown as the nucleosome occupancy at -1000bp. Similarly, average nucleosome occupancy level per site was calculated from -900 to -800bp, -800 to -700bp, -700 to

-600bp, -600 to -500bp, -500 to -400bp, -400 to -300bp, -300 to -200bp, -200 to -150bp, -150 to -100bp, -100 to -50bp and -50bp to +10bp regions.

Transcription start site (TSS) data for all yeast genes were obtained from Lu and Lin, 2019. Closest TSS site for each gene was obtained, and a spread of potential TSS sites for all genes were calculated. The mRNA synthesis rates and decay rates were obtained from Sun *et al*., 2012. Genome-wide mRNA secondary structure in yeast was obtained from Kertesz *et al.*, 2010. The mRNA half-life data and the protein half-life data was obtained from Geisberg *et al.*, 2014 and Belle *et al.*, 2006 respectively.

Genome-wide histone modification data for yeast was obtained from Pokholok *et al.*, 2006 and all modifications were mapped to Gene body, promoter region, and transcription factor binding sites. Histone binding dynamics was obtained from Dion et al., 2007. Data on three-dimensional model of yeast genome was obtained from Duan and Andronescu *et al*., 2010. Number of intra- and Inter- chromosomal contacts for all genes and promoter regions were quantified. tRNA adaptation index for all genes were calculated following the method of Tuller *et al*., 2010. For all molecular factors, the value of a factor in a biological process was obtained by calculating the median values of the factor for all genes in the process.

The classification for SAGA or TFIID dependence of genes for their expression were obtained from Huisinga and Pugh (2004). In addition, co-activator redundant or TFIID dependent classification of genes were used from Donczew *et al.*, 2020. The binding activity of several broadly acting TFs such as TBP, ABF1 and RAP1 were obtained from van Werven *et al.*, 2009, Lickwar *et al*., 2012 and de Jonge *et al*., 2020 respectively. Binding patterns of chromatin remodelers were obtained from Yen *et al.*, 2012; Zentner and Henikoff, 2013 and Ramachandran *et al.*, 2015.

Analysis of transcription factor binding

The list of transcription factors for all yeast genes were obtained from Yeastract database (http://www.yeastract.com/) (Teixeira *et al.*, 2018) and for a gene, only those TFs for which experimental evidence for DNA binding has been obtained or the knockout of TF has been experimentally shown to impact expression of the gene (Teixeira *et al.*, 2018) were considered. In addition, the binding sites of all TFs to promoter regions of the target genes were searched and mapped using the consensus motif sequences for TFs obtained from YeTFaSCo database (de Boer and Hughes, 2012). All position weighted matrices for all motifs were obtained and all possible combinations of bases were considered. Positions of all such motif sequences of all regulatory TFs of a gene were identified in the promoter region (ranging from -1000bp to +10bp of the start codon) after allowing for maximum two mutations in the consensus sequence. Overlaps between TF binding sites in promoter regions were mapped. The list of co-operatively binding TFs in yeast were obtained from Yang *et al*., 2010 and Chen *et al*., 2012. If the noise value for a TF was more than 0.5 (DM-value), the TF was considered to be a high-noise TF. Similarly, a TF showing a noise value of less than -0.5 (DM-value) was considered to be a low-noise TF.

Gene-transcription factor (TF) and TF-TF expression correlation analysis

Gene expression data measured through RNA sequencing from Dhar *et al*., 2019 (NCBI GEO dataset id 104343) was used to calculate all pairwise gene-TF and all pairwise TF-TF (of a gene) expression correlations. Significant positive correlation ($p<0.05$) between a gene and a regulatory TF indicated that the TF acted as an activator for the gene since the expression of the gene increased with increase in expression of the TF. Similarly, Significant negative correlation between a gene and its regulatory TF indicated that the TF acted as a repressor for the gene. The value of correlation coefficient between a gene and a TF (if significant) was taken as the response correlation and the slope of the line was considered as the strength of regulation of the TF.

In addition, pairwise expression correlations between all TFs of a gene were calculated. If a TF showed significant positive correlation with at least three other TFs, the TF was considered to be a positively correlated (co-expressing) TF. Similarly, if a TF showed significant negative correlation with at least three other TFs, the TF was considered to be a negatively correlated TF.

Modeling and stochastic simulation of transcription factor-DNA binding process

The dynamics of TF binding to DNA was captured using a two-state model with consideration of rapid binding and unbinding of TF to DNA. The binding-unbinding of TF to DNA was considered to be a Poisson process and thus, the time intervals between two successive bindings (or two successive unbindings) were exponentially distributed. The dynamics of co-operative binding and competitive binding of TFs was compared to the dynamics of the single TF regulation system. The co-operative binding was modeled by co-operative interactions between two TFs that required binding of two co-operative TFs together to the promoter binding site for transcription to take place. Competitive binding was modeled using competition between three TFs. The binding rate K_a and dissociation rate K_d were considered to be the same for all TFs across three scenarios to make the comparisons mathematically equivalent. As the concentrations of TFs can impact the chances of binding, the concentration of TF in single TF binding scenario was considered to be same as the concentration of each of the co-operatively binding TFs. Further, the concentration of the TF in single TF binding scenario was considered to be equal to the sum of the concentrations of three TFs in case of competitive binding scenario. Binding of TF led to ON state and resulted in transcription leading to accumulation of protein (Supplementary fig. S5A).

The dynamics of transcription and translation were modeled using the following equations.

$Changes\ in\ mRNA\ conc.\ over\ time$: $\frac{d[mRNA]}{dt} = \beta_m - \alpha_m \times [\mathrm{mRNA}]$

where β_m denoted the transcription rate per unit time (or burst size) and α_m denoted the removal rate of mRNA due to degradation and dilution.

Similarly,

$Changes\ in\ protein\ conc.\ over\ time$: $\frac{d[P]}{dt} = \beta_p \times [mRNA] - \alpha_p[P]$

where β_p denoted the protein production rate from mRNA and α_p denoted the protein removal rate. The parameters β_p , α_m , α_p were considered to be same across all cases of TF binding. In case of co-operative TF binding, $\beta_{m,coop}$ was assumed to be the same as the $\beta_{m,single.}$ In case of competitive TF binding, the production rates varied between three TFs, with $\beta_{m1} = 1.4\times\beta_{m,single}$, $\beta_{m2} = \beta_{m,single}$, and $\beta_{m3} = 0.7\times\beta_{m,single}$.

Stochastic simulations were performed using Gillespie's algorithm (Gillespie, 1977) to decipher the dynamics of TF-DNA binding in all three scenarios and to investigate the impact of co-operative and competitive TF binding on noise. The behavior of the system was tracked at small discrete time intervals Δt from the initial time point t. These resulted in observations at n time points t, t+ Δt, t+2×Δt, …, t+n×Δt. Any event of binding or unbinding occurring within a time interval was noted and resulted in changes in transcription rate which eventually led to a change in protein concentration (Supplementary fig. 5B). Binding of TFs led to transcription and increase in mRNA and protein concentration according to the above equations. As the time interval Δt was considered to be small, the equations modeling the behavior of the systems was simplified as

$$[mRNA]_{t+\Delta t} = [mRNA]_t + (\beta_m - \alpha_m \times [mRNA]_t) \times \Delta t$$

and

$$[P]_{t+\Delta t} = [P]_t + (\beta_p \times [mRNA]_t - \alpha_p \times [P]_t) \times \Delta t$$

β_m , α_m , β_p and α_p were expressed in the units of Δt for further simplification of these equations. The dynamics of transcription, variation in the mRNA concentration and variation in protein concentration with time were modeled across 10,000 cells (Supplementary fig. S5B). Noise was expressed as coefficient of variation (CV) from the calculation of mean and standard deviation in the protein level across these 10,000 cells and across all individual time points.

Statistical analysis

For all correlation analysis, Pearson's correlation coefficient was calculated. For all statistical comparisons between high- and low- noise genes Mann Whitney U test was performed.

Linear regression analysis

To estimate the percentage of variation explained by each factor, linear regression analysis was performed on our integrated dataset. The dataset was first scaled using z-score standardization. Next, to estimate the impact of each factor, the data was divided into training and test sets, the training set consisted of 80% of the whole dataset whereas the test set contained the remaining 20% of the dataset. Linear regression was performed on the training set to obtain the percentage of variation explained. The linear regression model thus built was applied on test dataset to obtain predicted R^2 value and RMSE values. This process of training and prediction was repeated 1000 times to obtain mean and standard deviation values for percentage of variation explained as well as to get predicted R^2 and RMSE values.

For estimating the predictive power of a class of factors on noise, a variable selection method using Ridge and Lasso regression was performed to minimize the problems of multi-collinearity and overfitting. Ridge regression was performed with the R package 'ridge' (Cule *et al*., 2020), appropriate number of principal components was chosen and factors showing significant effect on noise were identified. Lasso regression was performed using the R package 'glmnet' (Friedman *et al*., 2010) and factors with non-zero coefficients were identified as factors influencing noise. We then chose the most important factors through stepwise addition and removal of factors using stepwise regression where Akaike Information Criterion (AIC) of the fitted models were minimized. This was done using the R package 'olsrr' (https://github.com/rsquaredacademy/olsrr). The factors in the model with lowest AIC values were selected for further analysis.

In the next step, linear regression with the selected factors was performed on the training set to obtain the percentage of variation explained. The linear regression model obtained was then applied on the test data to obtain predicted values along with predicted R^2 values. This process of training and prediction was repeated 1000 times to obtain mean and standard deviation values for percentage of variation explained, predicted R^2 values. Further, random forest models were also built using the selected factors using the R package 'randomForestSRC' both with and without missing value imputations. These also resulted in percentage of variation explained and predicted R^2 values.

In addition to the original dataset, two filtered datasets were created with reduced number of variables – one filtered on correlation and another filtered on impact. The first filtered dataset was created by removing factors that did not show significant correlation ($p<0.05$) with process noise. To create the second filtered set, first, the impact of all individual factors on noise were obtained by linear regression as described above. Only the factors that showed significant impact (explained at least 5% of the noise variation or had predicted R^2 of more than 0.05) were retained in the filtered set. Linear regression was performed on these datasets as described above to obtain percentage of variation explained and R^2 values for prediction. The analysis showing best results for percentage of variation explained and predicted R^2 were reported.

References

Arnoldini, M. *et al.* Bistable expression of virulence genes in salmonella leads to the formation of an antibiotic-tolerant subpopulation. *PLoS Biol.* **12**, (2014).

Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000)

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as phenotypic switch. *Science* **305**, 1622–1625 (2004)

Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38**, 636–643 (2006).

Baudrimont, A., Jaquet, V., Wallerich, S., Voegeli, S. & Becskei, A. Contribution of RNA Degradation to Intrinsic and Extrinsic Noise in Gene Expression. *Cell Rep.* **26**, 3752-3761.e5 (2019).

Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA* **103**, 13004-9. (2006)

Bergmiller, T. *et al.* Biased partitioning of the multidrug efflux pump AcrAB-TolC underlies long-lived phenotypic heterogeneity. *Science* **356**, 311–315 (2017).

Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).

Burga, A., Casanueva, M. O. & Lehner, B. Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* **480**, 250–253 (2011).

Burger, A., Walczak, A. M. & Wolynes, P. G. Abduction and asylum in the lives of transcription factors. *Proc. Natl. Acad. Sci. USA* **107**:4016-21 (2010)

Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019)

Chen, M.J. *et al.* De novo motif discovery facilitates identification of interactions between transcription factors in Saccharomyces cerevisiae. *Bioinformatics* **28**, 701-8. (2012)

Chen, X. & Zhang, J. The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. *Cell Syst.* **2**, 347–54 (2016).

Cherry, J. M. *et al.* Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res.* **40**, (2012)

Choi, J. K. & Kim, Y. J. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat. Genet.* **41**, 498–503 (2009)

Cule, E., Moritz, S., & Frankowski, D. ridge: Ridge Regression with Automatic Selection of the Penalty Parameter. R package version 2.5, https://CRAN.R-project.org/package=ridge. (2020)

das Neves, R. P. *et al.* Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol.* **8**, e1000560 (2010).

Das, D., Dey, S., Brewster, R. C. & Choubey, S. Effect of transcription factor resource sharing on gene expression noise. *PLoS Comput. Biol.* **13**, e1005491 (2017).

de Boer, C. G. & Hughes, T. R. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res*. **40**, D169-79 (2012).

de Jonge, W.J., Brok, M., Lijnzaad, P., Kemmeren, P. and Holstege, F.C. Genome-wide off-rates reveal how DNA binding dynamics shape transcription factor function. *Mol Syst Biol*. **16**: e9885. (2020)

Dewachter, L., Fauvart, M. & Michiels, J. Bacterial Heterogeneity and Antibiotic Survival: Understanding and Combatting Persistence and Heteroresistance. *Molecular Cell* **76**, 255–267 (2019)

Dhar, R., Missarova, A. M., Lehner, B. & Carey, L. B. Single cell functional genomics reveals the importance of mitochondria in cell-to-cell phenotypic variation. *Elife* **8**, e38904 (2019)

Dickinson, M. E. *et al*. High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).

Dion, M. F., Kaplan, T., Kim, M., Buratowski, S., Friedman, N., Rando, O.J.. Dynamics of replication-independent histone turnover in budding yeast. *Science* **315**, 1405-8. (2007)

Donczew, R., Warfield, L., Pacheco, D., Erijman, A. and Hahn, S. Two roles for the yeast transcription coactivator SAGA and a set of genes redundantly regulated by TFIID and SAGA. *Elife* **9**, e50109. (2020)

Donovan, B.T., Huynh, A., Ball, D.A., Patel, H.P., Poirier, M.G., Larson, D.R., Ferguson, M.L., and Lenstra, T.L. Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *EMBO J*. **38** (2019).

Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).

Eldar, A. *et al.* Partial penetrance facilitates developmental evolution in bacteria. *Nature* **460**, 510–514 (2009).

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).

Engl, C., Jovanovic, G., Brackston, R.D., Kotta-Loizou, I., & Buck, M. The route to transcription initiation determines the mode of transcriptional bursting in E. coli. *Nat. Commun.* **11** (2020).

Farquhar, K. S. *et al.* Role of network-mediated stochasticity in mammalian drug resistance. *Nat. Commun.* **10**, 2766 (2019)

Faure, A. J., Schmiedel, J. M. & Lehner, B. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst.* **5**, 471-484.e4 (2017)

Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010)

Geisberg, J.V., Moqtaderi, Z., Fan, X., Ozsolak, F., Struhl, K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156**:812-24. (2014)

Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977)

Gupta, P. B., Pastushenko, I., Skibinski, A., Blanpain, C. & Kuperwasser, C. Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance. *Cell Stem Cell* **24**, 65–78 (2019)

Hammerlindl, H. & Schaider, H. Tumor cell-intrinsic phenotypic plasticity facilitates adaptive cellular

reprogramming driving acquired drug resistance. *J. Cell Commun. Signal.* **12**, 133–141 (2018)

Hornung, G. *et al.* Noise-mean relationship in mutated promoters. *Genome Res.* **22**, 2409–17 (2012).

Huisinga, K. L. and Pugh, B. F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. *Mol Cell.* **13**, 573-85. (2004)

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., Segal, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-7. (2010)

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

Levin, B. R. & Rozen, D. E. Non-inherited antibiotic resistance. *Nature Reviews Microbiology* **4**, 556–562 (2006).

Levin-Reisman, I. *et al.* Antibiotic tolerance facilitates the evolution of resistance. *Science* **355**, 826–830 (2017).

Lickwar, C.R., Mueller, F., Hanlon, S.E., McNally, J.G. and Lieb, J.D. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251-5. (2012)

Lu, Z. & Lin, Z. Pervasive and dynamic transcription initiation in Saccharomyces cerevisiae. *Genome Res.* **29**, 1198–1210 (2019)

Mcadams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 814–819 (1997)

Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013)

Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).

Nguyen, A., Yoshida, M., Goodarzi, H. & Tavazoie, S. F. Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. *Nat. Commun.* **7**, 11246 (2016)

Nicolas, D., Zoller, B., Suter, D. M. & Naef, F. Modulation of transcriptional burst frequency by histone acetylation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7153–7158 (2018).

Oberbeckmann, E. *et al.* Absolute nucleosome occupancy map for the *Saccharomyces cerevisiae* genome. *Genome Res.* **29**, 1996–2009 (2019).

Page, R. & Peti, W. Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nat. Chem. Biol.* **12**, 208–214 (2016).

Pisco, A. O. *et al.* Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nat. Commun.* **4**, 2467 (2013)

Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005)

Raj, A., Rifkin, S. A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–8 (2010).

Ramachandran, S., Zentner, G.E., and Henikoff, S. Asymmetric nucleosomes flank promoters in the budding yeast genome. *Genome Res.* **25**, 381-90. (2015)

Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science*. **304**, 1811–1814 (2004)

Ravarani, C. N. J., Chalancon, G., Breker, M., De Groot, N. S. & Babu, M. M. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nat. Commun.* **7**, 10417 (2016).

Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).

Rotem, E. *et al.* Regulation of phenotypic variability by a threshold-based mechanism underlies bacterial persistence. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12541–6 (2010).

Salari, R. *et al.* Teasing Apart Translational and Transcriptional Components of Stochastic Variations in eukaryotic Gene Expression. *PLoS Comput. Biol.* **8**, e1002644 (2012)

Sanchez, A., Garcia, H. G., Jones, D., Phillips, R. & Kondev, J. Effect of Promoter Architecture on the Cell-to-Cell Variability in Gene Expression. *PLoS Comput. Biol.* **7**, e1001100 (2011).

Sanchez, A., Choubey, S. & Kondev, J. Regulation of Noise in Gene Expression. *Annu. Rev. Biophys.* **42**, 469–491 (2013).

Silander, O. K. *et al.* A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli. *PLoS Genet.* **8**, e1002443 (2012).

Sharma, A. *et al.* Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors. *Cell Rep.* **29**, 2164-2174.e5 (2019)

Sharon, E. *et al.* Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res.* **24**, 1698–1706 (2014)

Sheng, S. *et al.* Tackling tumor heterogeneity and phenotypic plasticity in cancer precision medicine: our experience and a literature review. *Cancer Metastasis Rev.* **37**, 655–663 (2018)

Small, E. C., Xi, L., Wang, J. P., Widom, J. & Licht, J. D. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2462–E2471 (2014)

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K.C., Seizl, M., Tresch, A., and Cramer, P. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.* **22**,1350-9. (2012)

Taeubner, J. *et al.* Penetrance and Expressivity in Inherited Cancer Predisposing Syndromes. Trends in Cancer **4**, 718–728 (2018)

Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science.* **329**, 533–538 (2010).

Teixeira, M. C. *et al.* YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Res*. **46**, D348–D353 (2018).

Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* **38**, 830–834 (2006)

Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–1091 (2008).

Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010)

van Werven, F.J., van Teeffelen, H.A., Holstege, F.C. and Timmers, H.T. Distinct promoter dynamics of the basal transcription factor TBP across the yeast genome. *Nat Struct Mol Biol.* **16**, 1043-8. (2009)

Wakamoto, Y. *et al.* Dynamic persistence of antibiotic-stressed mycobacteria. *Science* **339**, 91–95 (2013).

Wang, Y., Ni, T., Wang, W., and Liu, F. (2019). Gene transcription in bursting: a unified mode for realizing accuracy and stochasticity. *Biol. Rev.* **94**, 248–258.

Weinberger, L. *et al.* Expression Noise and Acetylation Profiles Distinguish HDAC Functions. *Mol. Cell* **47**, 193–202 (2012).

Wu, S. *et al.* Independent regulation of gene expression level and noise by histone modifications. *PLoS Comput. Biol.* **13**, e1005585 (2017).

Yang, Y., Zhang, Z., Li, Y., Zhu, X. G. and Liu, Q. Identifying cooperative transcription factors by combining ChIP-chip data and knockout data. *Cell Research* **20**, 1276–1278 (2010)

Yen, K., Vinayachandran, V., Batta, K., Koerber, R.T., and Pugh, B.F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* **149**, 1461-73 (2012).

Zentner, G.E. and Henikoff S. Mot1 redistributes TBP from TATA-containing to TATA-less promoters. *Mol Cell Biol.* **33**:4996-5004. (2013)

Zoller, B., Nicolas, D., Molina, N., and Naef, F. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol. Syst. Biol.* **11**, 823. (2015)

Availability of data and materials

The datasets analysed during the current study are available in the NCBI-GEO repository.

Competing interests

The authors declare no competing interests.

Funding

Work in the lab of RD was funded by an ISIRD grant from IIT Kharagpur and the Early career research (ECR) grant (ECR/2017/002328) from Science and Engineering research Board (SERB), India.

Authors' contributions

RD conceived the study, RD, LP, and SP performed data analysis, RD wrote the manuscript with inputs from LP and SP. All authors read and approved the manuscript.

Acknowledgments

We are highly grateful to Dr. Ben Lehner for his insightful and critical comments on the first draft of this manuscript.