M.Sc. Thesis

Noise in protein abundance obeys a general scaling law over multiple genes and growth conditions.

Arren Bar-Even

M.Sc. Advisor
Dr. Yitzhak Pilpel

Department of molecular genetics,
Weizmann Institute of Science

Nov 30, 2005

Index

ABSTRACT

In the last few years, a number of studies have shown that genetically identical cells growing under identical conditions still vary greatly in their internal protein concentrations. However, most studies have focused on a limited number of genes and got inconsistent results and varied conclusions. As a step in the direction of deciphering general patterns of genetic noise we measure the cell-to-cell variation of several tens of proteins from a library of single-reporter GFP translation fusions in *S. cerevisiae.* These represent four strongly co-expressed sets of genes (modules), whose mean expression levels span several orders of magnitude (tens of copies to hundreds of thousands). All genes were measured under eleven different conditions, at six consecutive time-points.

We show that at low average expression levels, the protein abundance distributions were approximately normal, while at higher averages, they were better approximated as log-normal. As a general experimental set-up validation step, we found that the mean abundance of the proteins is highly depended on their module affiliation, as expected. We show a strong correlation between variability and average expression level. Strikingly, in all gene modules and in a broad range of expression levels, the standard deviation is roughly proportional to the square root of the average. We suggest how this effect can be explained by fluctuations in the mRNA abundance. By analyzing the deviation from this general shape, we find that the genes cluster in a module-specific manner. In particular, for a given average, stress genes appear to be significantly noisier than other genes tested. We found that genetic noise has a bottom border, and hypothesize that this border originates from the un-equal division of the yeast cells. Overall, we show, for the first time, several general noise patterns, which affect a broad range of genes and pathways.

INTRODUCTION

Protein, mRNA and DNA molecules are present, in living cells, in a relatively low abundance(1-11). The consequences of that are significant stochastic effects that are responsible to the variation observed between cells in an isogenic population. Some organisms exploit noise in order to introduce diversity into a population, as in the cases of phenotypic variation in the lambda phase lysis-lysogenic switch(12), the mammalian olfactory neuronal receptor choice(13) or the DNA inversion mechanism in bacteria(14). However, in most cases fluctuations in protein abundance present a problem the cell must cope with. For example, stability against genetic noise is essential in cellular network controlling differentiation in the embryo development(15). In other cases stochastic fluctuations can lead to low fidelity in cellular behavior(6).

There are several commonly used definitions of noise in protein abundance; all are functions of the population mean protein abundance, μ and its standard deviation, σ: The coefficient of variation (CV): $\eta = \sigma/\mu$; the normalized variance: $\eta^2=\sigma^2/\mu^2$; and the noise strength: $\nu=\sigma^2/\mu$. The normalized variance is more suitable when one wants to add-up several stochastic sources, affecting the same downstream protein. Moreover, both the normalized variance and the CV are unit free measurements and hence are not affected by the measurements scale. The noise strength, also known as the Fano factor, is convenient for size-independent comparison for Poisson processes(7). Those processes are characterized in $\sigma^2/\mu=1$, and therefore the noise strength measures deviations from Poisson behavior(16).

The variability in the abundance of a specific protein can be attributed to two different sources. The protein abundance is influenced by many up-stream cellular entities, such as amounts and concentrations of regulatory proteins, ribosomes, polymerases and most important the mRNA abundance translated to create the protein. Fluctuation in their amount will propagate down-stream to the protein abundance level. Those contributions are defined as the extrinsic source(7, 17). However, even in the hypothetical situation in which the abundances of all the up-stream elements were equal in all the population cells, the random nature of the microscopic events, governing the protein production and elimination reactions, will create fluctuations in

its abundance. This source of noise is termed intrinsic. However, the exact distinction between the intrinsic and the extrinsic noise depends on the experimental set-up and on the mathematical framework used. We will use the strict definition of intrinsic noise – noise originating from the random births and deats of individual protein molecules. Hence, all other sources of noise, including fluctuation of mRNA abundance will be regarded as the extrinsic noise of the protein abundance(7). Extrinsic noise itself can be viewed as composed of three parts: global noise, influencing all cellular proteins, such as cell division and ribosomes or polymerases abundance; module related noise, affecting all proteins belonging to a single regulated module, which can originates from fluctuations in the common regulators; and gene specific (individual) extrinsic noise which is created, for example, by special chromatin properties near the gene or by the mRNA level specific for the genes. A summary of the discussed noise types is given in figure 1.

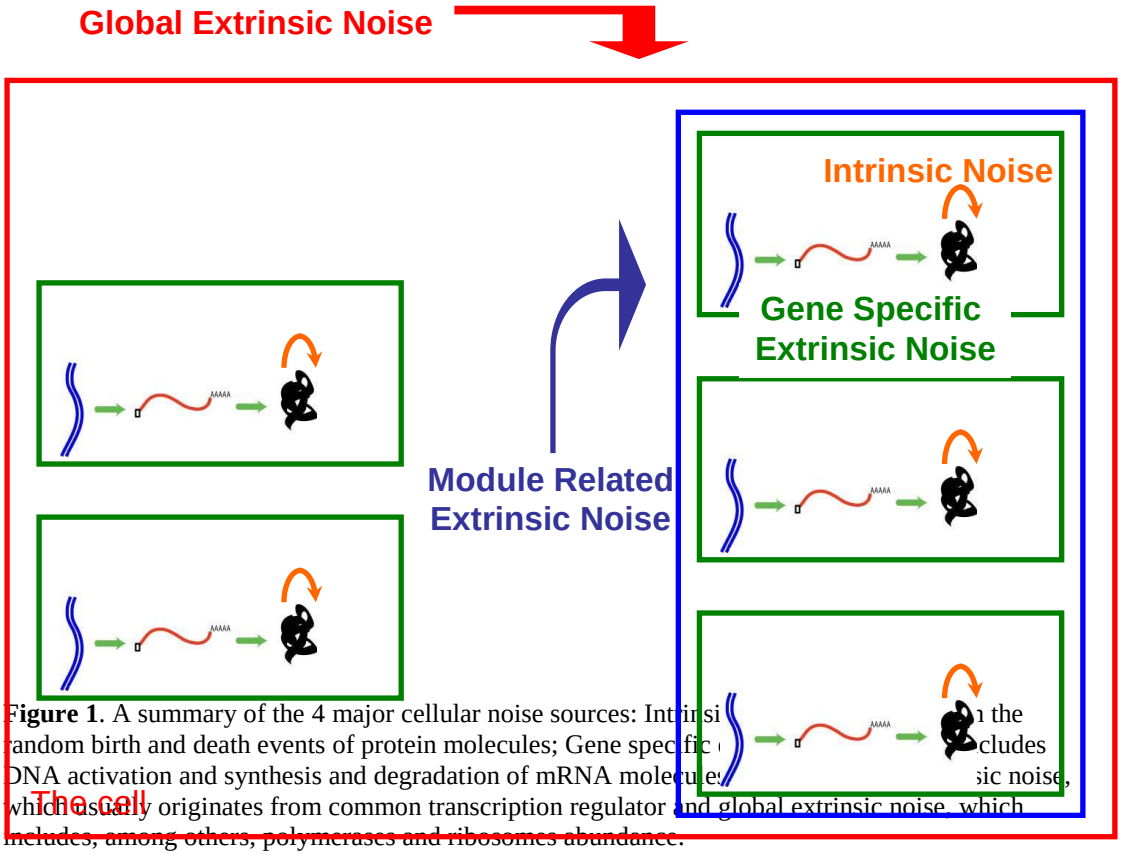

Figure 1. A summary of the 4 major cellular noise sources: Intrinsi[illegible] the random birth and death events of protein molecules; Gene specific [illegible]cludes DNA activation and synthesis and degradation of mRNA molecules[illegible]sic noise, which usually originates from common transcription regulator and global extrinsic noise, which includes, among others, polymerases and ribosomes abundance.

A mathematical theory of noise

Several mathematical formalisms have been used to model noise pattern of cellular proteins. Of those, the most comprehensive, yet simple, one was crated by Johan

Paulsson, which approximates both the intrinsic and extrinsic fluctuations in the protein abundance(7). Consider a system in which one chemical species X_1 affects the rate of production, or elimination, of other chemical species X_2, but not the other way around, such that:

$$n_1 \xrightarrow{R_1^{\pm}(n_1)} n_1 + 1 \quad , \quad n_2 \xrightarrow{R_2^{\pm}(n_1, n_2)} n_2 + 1$$

where n_1 and n_2 are the abundances of X_1 and X_2, respectively.

And to quote:

"…X1 provides the randomly fluctuating environment for X2, as mRNA fluctuations randomize protein synthesis. To collectively approximate all such processes I use the Ω-expansion where the first- and second-order terms reproduce the macroscopic rate equations and the fluctuation-dissipation theorem respectively. The latter is then interpreted in terms of the logarithmic gains $H_{ij}=\partial\ln(R_i^-/R_i^+)/\partial\ln(n_j)$ that measure how the balance between production and elimination of Xi is affected by Xj and can often be estimated directly from the reaction rates. For the process described above, using σ_i for standard deviations, μ_i for averages and τ_i for average lifetimes, stationary fluctuations around a stable fixed point follow

$$\frac{\sigma_2^2}{\mu_2^2} = \frac{1}{\mu_2 H_{22}} + \frac{\sigma_1^2}{\mu_1^2}\frac{H_{21}^2}{H_{22}^2}\frac{H_{22}/\tau_2}{H_{11}/\tau_1 + H_{22}/\tau_2}$$

Where $\frac{\sigma_1^2}{\mu_1^2} = \frac{1}{\mu_1 H_{11}}$ "

The first term represents the intrinsic noise, which depends on the mean abundance, while the second term corresponds to the extrinsic noise. The latter is composed of three elements. First is the noise in the upstream species, σ_1^2/μ_1^2 (which equals $1/\mu_1 H_{11}$ in case X1 does not have its own extrinsic source). The susceptibility factor, H_{21}^2/H_{22}^2, is a measurement of the sensitivity of X_2 to the changes in abundance of X_1. This term is related to the slope of the graph connecting the mean abundance of X_1 to the production or elimination rate of X_2. High susceptibility (high slope) means that small changes in the abundance of X_1 will cause large changes in the abundance of X_2, therefore creating large fluctuations in X_2 in response to small fluctuations in X_1. The last element, $(H_{22}/\tau_2)/(H_{11}/\tau_1 + H_{22}/\tau_2)$, is the time averaging effect, which is taking into account consecutive changes in X_1 that cancel out due to different time scale of the two species abundance changes. The use of normalized variance was advantageous because extrinsic noise from parallel sources make super-imposable contributions to it. It is important to note that the intrinsic noise, as Paulsson defined it, is indeed only protein-related. Any fluctuations in the mRNA abundance will result in contribution to the extrinsic part of the equation.

Deciphering noise patterns – A short overview of experimental set-ups

Several methods have been used to track the origin of fluctuations and to investigate how they depend on average expression levels. One strategy is to measure the standard deviation in protein abundance as a function of the it's mean and of the transcription and translation rates(16, 18-21). Low-copy fluctuating proteins typically display a particular scaling behavior where the standard deviation is proportional to the square root of the average. The noise strength, in this case, does not scale both with the transcription and translation efficiencies. However, if the noise strength only scaled with translation but not with transcription, it has been argued (7, 16, 18) that noise probably comes from having few mRNA copies, not from having few protein copies. Applying this method to *Bacillus subtilis* suggested that most noise came from transcription and mRNA degradation(16), i.e., from low-copy mRNA fluctuations. Applications to *Saccharomyces cerevisiae*(21, 22) instead suggested that gene specific pre-transcription processes dominate under most conditions.

Another strategy is to measure the simultaneous expression of two identically regulated reporter genes(17, 19, 20, 23, 24). Each protein then has its own set of genes and mRNAs, but share both global and pathway-specific factors with the other protein. Because identical reporters should be equally susceptible to any fluctuations, this elegant method makes interpretation less model dependent: as long as the two reporter systems do not affect each other, the noise contribution from the shared environment should equal the covariance between the reporters. Noise specific to each reporter consists of intrinsic and gene-specific extrinsic contributions, while the shared noise is comprise of module-related and global contributions. An application to *Escherichia coli*(17) showed very little noise from global factors, at least for the genes investigated under the conditions used. Some noise was specific to each reporter, while the shared noise was largely explained by a pathway-specific repressor. Other works(19, 24) in the same organism demonstrated that global factors have a significant contribution to the overall noise. Most importantly, the cell-cycle was shown to have a dominant contribution to the fluctuation in protein abundance(19). Applying the same method to genes in *S. cerevisiae*(20, 23) showed that most noise originates from global factors, rather than pathway or gene-specific ones.

In the following sections we will elaborate on the experiments done and the conclusions made by their analysis.

Transcription and translation sources of noise in the prokaryote B. subtilis

The two basic process, directly influencing the protein abundance, are the transcription and translation (as well as mRNA and protein degradation). Oudenaarden and his group investigated the effect of those processes on the fluctuations of protein levels(16, 18). They modeled a simple system, containing mRNA molecules that are synthesized at rate of k_R from a template DNA strand, protein that is translated at a rate of k_P of each mRNA molecule and mRNA and protein degradation rates of γ_R and γ_P, respectively. They used a Master Equation technique to model this system and found out that in steady state the mean protein level is $\mu= k_R b/\gamma_P$, while the noise, given as the noise strength, is $\sigma^2/\mu=(b/(1+c))+1$, where $b=k_P/\gamma_R$ is the average number of proteins produced per mRNA transcript and $c=\gamma_P/\gamma_R$ is the ratio of mRNA to protein lifetimes. Generally $c<<1$ and therefore $\sigma^2/\mu\sim b+1$. Hence they conclude that noise is affected only from translation rate and is indifferent to the transcription rate.

In order to validate those conclusions they integrate GFP into the chromosome of B. subtilis under the regulation of the LAC operon and changed both transcriptional and translational efficiencies. The transcription efficiency was perturbed by using different concentration of the IPTG inducer and by creating point mutations in the promoter. Translational efficiency was disrupted by generating point mutations in the ribosome binding site (RBS) or in the initiation codon of the GFP. Their results confirmed that transcriptional efficiency had only minor effect on the noise strength while translational efficiency changes it linearly.

However, these results are misleading(7). The burst term b does not come from the randomness to translation, but from the fluctuation in mRNA abundance. Eliminating those fluctuations will abolish that term. Looking at the normalized variance, and implementing Paulsson's models, under the same assumptions, we get $\sigma_p^2/\mu_p^2=1/\mu_p+(\tau_R/\tau_P)/\mu_R=1/\mu_p+k_P\tau_R/\mu_p=1/\mu_p+b/\mu_p$, where τ_R and τ_P are the averaged life

time of mRNA and protein molecules, respectively. Therefore the b term indeed represents the extrinsic source of noise. Plotting σ_p^2/μ_p^2 as a function of $1/\mu_p$ gave, as expected, a straight line with a small displacement.

The conclusion that translation rate, but not transcription rate, influences the noise strength has an interesting evolutionary consequence. If a protein is needed in a certain cellular abundance it could be produced by using poor efficiency transcription followed by high efficiency translation or by high efficiency transcription followed by poor efficiency translation. The second mechanism is much more expansive, but will lead to less noisy output, and the interplay between those two considerations will create a selective pressure. Essential genes or genes that work in a complex are among the genes that the cell might want the keep at low fluctuations level. Fraser et al. found that those genes are indeed characterized in relatively high ratio of translation rate to transcription rate(25).

Intrinsic and extrinsic noise measurement in E. coli using two reporters

In order to differentiate between the two main sources noise Elowitz et al. constructed strains of E. coli, incorporating CFP and YFP in the chromosome(17). In each strain the two reporter genes were controlled by identical promoters and were integrated at loci, equidistant from the origin of replication. Difference in the fluorescence of the two reporters, in the same cell, was considered to be a measure of the intrinsic noise, while the overall correlated fluorescence difference between different cells represent the extrinsic noise. This use of definitions makes a non-trivial distinction between system and environment(7) and is not consistent with the general scheme we presented. The measured intrinsic noise in these experiments includes also some of the gene specific extrinsic noise, as we defined above, most importantly the mRNA abundance.

The measured 'intrinsic' noise behaved almost like the total noise in the B. subtilis experiment(7): $\sigma_p^2/\mu_p^2 \sim 1/\mu_p + C$. However the extrinsic contribution was dominant in almost all cases, which contradict the former study. A possible explanation to this pattern is the noisy nature of the lacI repressor they used to control their reporters, which was incorporated into a plasmid or under oscillating control.

Unsurprisingly, the authors found out that strong induction lowered both the intrinsic and the extrinsic noise while oscillating regulator abundance increased protein fluctuation. One interesting finding was that a deletion of RecA, which acts to rescue stalled replication forks, doubles the noise level. Hence, it was suggested that increased noise may arise from transient copy number differences between different parts of the chromosome.

Transcription effect on noise strength is yeast

Blake et al. turn their focus to the eukaryotic S. cerevisiae yeast(22). The authors built an artificial genetic network, in which the repressor TetR is regulates by a galactose responsive promoter. The expression of GFP is under the control of the repressor TetR; a repression that was tuned by the concentration of the inducer ATc. Varying the concentration of glactose or ATc affects the expression rate of the reporter gene.

In contrast to noise strength in prokaryotic bacteria, which was insensitive to transcription efficiency, the noise strength in the yeast changed non-monotonically with transcription rate – having a peak at partial expression induction and decreasing both at low or full induction. This phenomenon was attributed to eukaryotic unique mechanisms, such as chromatin remodeling and the formation of pre-initiation complex. Additionally, translation rate has the largest effect on the noise strength also at partial induction. The reason for this pattern is probably the fact that translation efficiency just amplifies the upstream noise. Hence, the translation effect on noise strength is maximal when the upstream transcription process has its maximum effect on stochasticity.

Two reporters experiments in S. cerevisiae

Raser et al. used the two reporters experiment to explore the pattern of extrinsic and intrinsic noise in S. cerevisiae(23). There first important finding was that extrinsic noise dominate the total protein fluctuations. Intrinsic noise was only 2% to 20% of the overall noise. Moreover, in an attempt to differentiate between gene specific extrinsic noise and global extrinsic noise they attached the two reporters to the promoters of different genes, not even belonging to the same module (in one case they

used the promoters of PHO84 and GAL1 and in anther case the promoters were of the genes PHO84 and ADH1). The correlation between the two fluorescence level was very high (R^2 of 0.88 and 0.93, respectively). Therefore, they concluded that the majority of extrinsic noise is from a global origin.

The authors also found that although the extrinsic noise of two of the examined genes (GAL1 and PHO85) behave according to the model proposed by Thattai et al. (that is, transcription had no effect on intrinsic noise strength)(16, 18), one gene (PHO5) behaved differently. Decreasing PHO5 rate of expression more than doubled the intrinsic noise strength. In order to explain this unexpected pattern a model was created. The model took into consideration also the activation of the DNA, by chromatin remodeling, apart from the transcription and translation. Three different kinetic mechanisms were proposed to explain different behaviors of intrinsic noise, as shown in figure 2. In the first mechanism, which corresponds to the behavior of PHO5, the activation of the DNA is infrequent, but stable (that is slow inactivation). The noise strength, assuming this model, will increase both with transcription and translation efficiencies but will decrease with higher DNA activation rate. The second mechanism is characterized with slow and un-stable DNA activation. In this case noise strength will still increase with transcriptional and translational efficiencies, but will remain relatively unchanged by increased DNA activation rate. The last mechanism, which corresponds to that of Thattai et al, assumes high activation and in-activation rates. Hence, as compared to the more slow transcription event, those processes are averaged and do not contribute to the noise strength. As in Thattai et al model the only effect on the intrinsic noise strength will be that of changing the translation efficiency.

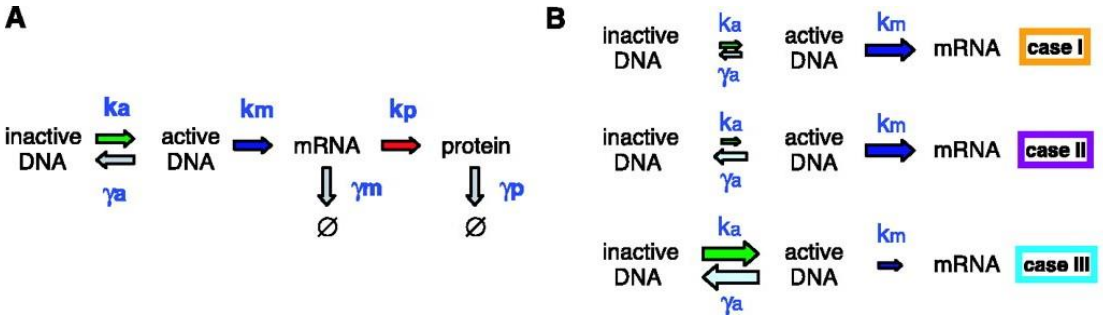

Figure 2. (taken from Raser et al). **A**. An extended model of protein production, including DNA activation and inactivation. **B**. Three different cases of relationships between the reactions rates can produced different intrinsic noise strength pattern, as described in the text.

In order to validate the first mechanism, the authors created mutations in the PHO5 upstream activating sequences UAS1 and UAS2 and deleted several component of the chromatin remodeling complexes. In all these cases an increase in the intrinsic noise strength was observed, as the model predicted. Damaging the TATA sequence, which is required for efficient transcription but dispensable for chromatin remodeling, resulted in an opposite trend of noise reduction, again as predicted.

Fluctuations in the gene regulation function, in E. coli

Rosenfeld et al. took a whole different approach of measuring cellular variation(19). They used a two reporter system, in which a repressor bound to YFP regulates the production of CFP. Before experiment start the production of the repressor is induced and at t=0 its production stopped, so it is diluted continually across cell divisions. The concentration of YFP slowly decreases, while CFP concentration rises along the progenies. The authors tried to calculate the mean function that describes the relation between the repressor (YFP) concentration and the YFP production rate, and the characteristic deviation of single cells from it.

It was found that the partition of YFP between the two daughter cells, upon cell division follows a binomial rule, therefore creating a source of noise that is proportional to the square root of the total abundance. The CFP production was strongly correlated to the cell cycle phase – cells just before division produce double the amount of CFP than expected by the amount of the repressor.

The origin of the remaining noise (deviation from the expected function) was tested by placing the two different reporters under the control of the same promoter, regulated by the above repressor. It was found that intrinsic noise capture only 20% of that remaining noise. Moreover, the authors found that intrinsic noise tends to fluctuate very rapidly, so it is being averaged out in the life period of a single cell. Extrinsic noise, which was much more dominant, was characterized in a time scale of cell cycle period, and therefore creates a real individuality between single cells. In addition, because noise here was a measure relative to the concentration of the repressor, the extrinsic noise was related to mostly global factors.

Noise propagation in a cascade

Differentiating between different sources of extrinsic noise is very important in order to reveal the dominant noise contributor in the cellular environment. Pedraza et al. used E. coli and incorporated in it a cascade that consists of the three elements(24). The repressor LacI was regulating the expression of the repressor tetR, which was attached to CFP. The reporter YFP, in turn, was regulated by the tetR. In addition the reporter RFP was placed under the regulation of constitutive, un-regulated, promoter. Both the repressions of LacI and of tetR were tunable using the inducers IPTG and ATC, respectively.

By measuring the auto- and cross-correlation of the fluorescence of the three reporters and by implementing Langevin modeling approach(26, 27) the authors separated both the intrinsic from the extrinsic noise and also the cascade propagated extrinsic noise from the global extrinsic noise. As in the other experimental work, intrinsic noise was found to constitute only a minor fraction of the overall noise. The propagated noise and the global noise, both have a dominant role in determining the reporters' fluctuation. The importance of the susceptibility factor was demonstrated. The contribution of the propagated noise raised more than 4 fold in the region in which the sensitivity of the regulated reporter to the repressor was maximal. In this region the propagated noise becomes the most important fluctuations source, shadowing all others.

The pattern of global noise was also examined. Because this source is affecting protein abundance both directly and indirectly through the upstream genes in the cascade, its contribution is not constant and strongly depends on the genetic circuit properties. Giving the effect of the susceptibility factor and that of the varying contribution of the global noise, even in a network where all components have low intrinsic noise, fluctuations can be substantial. However, those results should be taken with a grain of salt. The cascade built by the authors was placed on a plasmid, rather than incorporated into the chromosome. Therefore, the global extrinsic noise probably includes a significant factor of plasmid copy number fluctuations.

The effect of gene activation and chromosomal positioning on noise

One of the major problems in measuring the noise level in protein abundance is the high background fluorescence level, which prevents us from measuring the exact fluorescence level of low abundance reporter genes. In order to solve that problem Becskei et al. devised a genetic circuit for noise amplification(21). The potent transcriptional activator rtTA was placed under the control of the investigated promoter. When rtTA was bound to the inducer, doxycyline, it drives the expression of YFP. The amplification originated from two mechanisms. First, low abundant rtTA resulted in high abundant and measurable amount of YFP. In addition, embedding several rtTA binding sites in the YFP promoter led to increased cooperativity and therefore to high susceptibility factor. Hence, small fluctuations in the abundance of rtTA were multiplied in the elevated susceptibility factor to create significant and easily measured noise in the YFP.

Using this amplification devise, the authors measured noise in low abundant cell cycle proteins. Several of them were very noisy, which was considered to be an outcome of low mRNA copy. However, gene duplication did not dump that noise, as expected if noise were indeed intrinsic in origin. It was found out that the dominant noise source originated from random events of gene activation. These events are largely influenced from chromosomal positioning and indeed repositioning of noisy promoters on different chromosomal locations dumped the noise level considerably. Moreover, the fluctuations of reporters that were located in proximity on the chromosome were significantly correlated. The high fluctuating genes regulate other down stream genes and transmit their noise further in the cascade, eventually spreading it throughout the genetic network.

Noise sources decomposition in the yeast cell-fate decision system

Another recent attempt to decompose the extrinsic noise into module related and global compounds was done by Colman-Lerner et al(20). The authors used the C. cerevisiae (of **a** mating type) pheromone response pathway, responsive to varying concentration of α-factor, to decoupled noise from different origins. Two parallel two reporters systems were built. The first consists of YFP and CFP, both under the regulation of the same α-factor responsive promoter, while in the second the YFP was

still under the same regulation, but CFP was under the regulation of a constitutive promoter. Using that system three different noise types can be distinguish: the intrinsic noise, noise that originate from variability in the α-factor signal transmission pathway and global noise, related to the overall protein production capacity of the cell. Intrinsic noise was found to explain only about 2% of the overall noise, while the dominant source of protein abundance fluctuations was global in origin – about 75% of total noise. The latter finding also explain why there was a linear relationship between the standard deviation and the mean of the protein abundance distributions, and not root square dependency, as expected from Poisson processes, which dominate the intrinsic noise. Using a cell cycle arrest experiment, it was shown that variability in cell cycle position produce about half of the overall noise. The rest of the global noise was attributed to global factors, such as ribosomes, RNA polymerase II complexes or cellular energy level.

At lower α-factor concentration, pathway related fluctuation become the dominant factor to control noise level – 59% of the total noise. However, the authors discovered that at varying level of α-factor, the overall stochasticity remain rather constant, despite the increase in the pathway related noise. Hence, there seems to be a rather mysterious anti-correlation between the global and the pathway related noise factors. However, this buffering interpretation depends on the assumption that global noise affects the α-factor responsive promoter equally across varying pheromone level, an assumption that might be incorrect(28).

Motivation for the current work

By nature of the analysis, most studies have focused on a limited number of genes, many of which are highly expressed. Expression levels have also been tuned either by mutation or by varying the amounts of activators and repressors, sometimes using genes that are synthetically engineered. This is well motivated when analyzing any particular mechanism in detail, but because the results have varied from study to study and from condition to condition, it remains to be seen how generally the conclusions apply. As a step in that direction we here measure the cell-to-cell variation of 38 proteins from a library of single-reporter GFP translation fusions in *S. cerevisiae.* These represent four strongly co-expressed sets of genes (modules), whose mean

expression levels span several orders of magnitude (tens of copies to hundreds of thousands). As in previous studies, one purpose of the present work is to correlate noise and average expression levels. However, rather than tuning the expression of a particular gene by varying transcription or translation rates, we instead compare genes that naturally have different expression levels under the same external conditions.

We measured the noise level, of all genes, at six consecutive time points, in 11 different conditions. These extended datasets help us determines the relative effect of module affiliation and environmental conditions on noise, by evaluating the similarity of noise level patterns between genes from the same module. Also we wanted to check whether there are distinguished patterns, of noise in time, for each of the modules and conditions.

METHODS

Strains and growth conditions

We used 43 strains from the yeast GFP clone collection(29), bought from Invitrogen. The genotype of the parent haploid *S. cerevisiae* strain, ATCC 201388, is MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0. A construct containing GFP and a HIS marker was incorporated into the 3'UTR of the chosen genes, therefore creating a mature protein which has a GFP attached to its terminus.

We chose 43 genes belonging to 4 distinct modules: stress (12 genes), proteasome (10), ergosterol (10) and rRNA processing (11). The genes are given in table 1.

Internal Numbering	Gene	Yname	Module Affiliation	Internal Numbering	Gene	Yname	Module Affiliation
1	TPS2	YDR074W	Stress	23	ERG10	YPL028W	Ergosterol
2	HSP104	YLL026W	Stress	24	CIC1	YHR052W	rRNA Processing
3	HSP78	YDR258C	Stress	25	HSP42	YDR171W	Stress
4	SSE2	YBR169C	Stress	26	AAH1	YNL141W	rRNA Processing
5	NOC2	YOR206W	rRNA Processing	27	PRS4	YBL068W	rRNA Processing
6	GSY2	YLR258W	Stress	28	RPN8	YOR261C	Proteasome
7	ACS2	YLR153C	Ergosterol	29	PRE10	YOR362C	Proteasome
8	SSA4	YER103W	Stress	30	BRX1	YOL077C	rRNA Processing
9	ARX1	YDR101C	rRNA Processing	31	RPN12	YFR052W	Proteasome
10	PWP1	YLR196W	rRNA Processing	32	PRE4	YFR050C	Proteasome
11	URA7	YBL039C	rRNA Processing	33	PRE9	YGR135W	Proteasome
12	PGM2	YMR105C	Stress	34	PUP2	YGR253C	Proteasome
13	ERG5	YMR015C	Ergosterol	35	SCL1	YGL011C	Proteasome
14	RPN3	YER021W	Proteasome	36	HSP26	YBR072W	Stress
15	DBP3	YGL078C	rRNA Processing	37	APT1	YML022W	rRNA Processing
16	HXK1	YFR053C	Stress	38	CYB5	YNL111C	Ergosterol
17	TPS1	YBR126C	Stress	39	HSP12	YFL014W	Stress
18	ERG13	YML126C	Ergosterol	40	ERG1	YGR175C	Ergosterol
19	PRS1	YKL181W	rRNA Processing	41	ERG11	YHR007C	Ergosterol
20	RPN7	YPR108W	Proteasome	42	ERG6	YML008C	Ergosterol
21	RPN6	YDL097C	Proteasome	43	ERG3	YLR056W	Ergosterol
22	MVD1	YNR043W	Ergosterol				

Table 1: The genes chosen for the experiments.

Synthetic complete medium (SC) was prepared by dissolving 6.7g of Bacto YNB without amino acids (Difco), 20g of D-Glucose and 1.6g of full drop-out, containing all needed amino-acids and amino-bases, in 1L of DDW. Cells were inoculated to SC media from YPD agar plates. After growth in an incubator at 30^{o}c for about 6 hours they were diluted to fresh SC media and grown overnight in the Unimax1010 Incubator Shaker (Heidolph), at 30^{o}c, to reach an OD of ~0.2 before experiment start next morning. Along side the tested genes we have measured the background

fluorescence by using, in each condition, two control types. Both controls used the same *S. cerevisiae* strain, lacking the GFP. The first, termed 'C' was treated with the experimental conditions as the rest of the strains while the other, termed 'CC', was grown without experimental perturbation.

Experiments

All experiments were conducted on cells in 10ml media and OD of ~0.2 (unless otherwise mentioned). We have carried out 11 experiments. Eight of them are stress conditions, with various cellular effects, while the other three conditions are stress relaxation. The concentrations of the stress reagents were determine by conducting growth rate experiments using a wild range of concentrations. We pick up the concentrations with a mediocre effect on the growth rate. The list of conditions is given in table 2.

Stress Conditions			
Condition	*Concentration*	*Reagent cellular effect*	*Abbreviation*
Diamide	1.5mM	Oxidative agent	DMD
Hydrogen Peroxide (H_2O_2)	0.3mM	Oxidative agent	HO
Methyl Methane-Sulfonate	0.04%W/V	Mutagen, inferring with and causing damage to DNA	MMS
Heat Shock	$30^oc \rightarrow 37^oc$		HT
Dithiothreitol	4mM	Reducing agent	DTT
Clotrimazole	10μM	Inhibitor of the ergosterol pathway	CLT
Rapamycin	65ng/ml	Inhibitor of the TOR pathway	RPM
Ethanol	3%	Non-fermentable carbon source	ETN
Stress Relaxation Conditions			
Condition			*Abbreviation*
Nitrogen depletion relaxation			NTR
Stationary phase relaxation			STT
Glycerol growth relaxation by glucose addition			GLY

Table 2: Stress and stress relaxation experiments we conducted.

The exact implementations of each condition are given herein. **Diamide:** 52µl of 290mM diamide (Sigma-Aldrich) in DDW were added to 10ml of cells in SC media to reach a final concentration of 1.5mM. **Hydrogen peroxide (H_2O_2):** 34.2µl of 88mM H_2O_2 in DDW were added to reach a final concentration of 0.3mM. **Methyl methane-sulfonate:** 40.2µl of 10%W/V MMS (Sigma-Aldrich) in SC were added to reach a final concentration of 0.04%W/V. **Heat shock 30°c→37°c:** Temperature of the incubator was elevated to 37°c at the experiment start. **Dithiothreitol:** 44.2µl of 0.9M DTT (Sigma-Aldrich) in DDW were added to reach a final concentration of 4mM. **Clotrimazole:** 33.5µl of 3mM Clotrimazole (Sigma-Aldrich) in DMSO were added to reach a final concentration of 10µM. **Rapamycin:** 16.3µl of 40µg/ml rapamycin (Sigma-Aldrich) in DMSO were added to reach a final concentration of 65ng/ml. **Ethanol:** 280µl of pure ethanol were added to 10ml to reach a final 3% ethanol. **Nitrogen depletion relaxation:** SC nitrogen depleted medium was prepared by dissolving, in 1L of DDW, 1.7g of Bacto YNB without amino-acids and ammonium sulfate (Difco), 20g of D-Glucose and 10ml of a solution containing 1g of Uracil, 1g of Methionine, 5g of Leucine and 1/3g of ammonium sulfate per 500 ml. Cells were grown over night (at 30°c) in SC nitrogen deplete media and reach an OD~0.5 in the morning. At experiment start 128µl of 3M ammonium sulfate in DDW were added to 10ml of cells in media to reach a final concentration of 37.8mM, as in non-nitrogen-depleted SC media. **Stationary phase relaxation:** Cells were grown for two days (at 30°c) to reach a deep stationary phase. At experiment start 333µl of cells were diluted into 10ml, to reach a final OD of ~0.5. **Glycerol growth relaxation by glucose addition:** SC medium with Glycerol as the sole carbon source was prepared be dissolving 6.7g of Bacto YNB without amino-acids (Difco), 30ml of Glycerol and 1.6g of full drop-out containing all amino-acids and amino-bases in 1L of DDW. Cells were grown in the above media over night (at 30°c) and reach an OD of ~0.5 at morning. At experiment start 256µl of 4M D-Glucose in DDW were added to reach a concentration of 0.1M, as in SC media which have D-Glucose as the sole carbon source.

Flow cytometry measurements

Flow cytometry experiments were conducted using the Becton-Dickinson FACSAria machine. Six measurements were taken, in every 30 minutes, from experiment start: 0min, 30min, 60min, 90min, 120min and 150min. In every time point the following parameters were recorded for 100,000 cells: **1. Forward Scatter Width** (FSC-W), which corresponds to the time the cells moved in front of the laser beam. **2. Forward Scatter Area** (FSC-A), corresponds to the size and reflection properties of the cell. **3. Side Scatter Area** (SSC-A), corresponds to the granularity and other reflection properties of the cell. **4. The 515–545nm detector** of the blue laser (GFP-A), which measures the fluorescence in the wave length of the GFP emission. The experiments of each condition were divided to two days – in the first genes 1-21 were examined, while genes 22-43 were test in the second day. The two types of controls, 'C' and 'CC', were measured in each of these days anew.

Figure 3 demonstrates the temporal evolution, of three of the above parameters, of a certain gene in a specific condition.

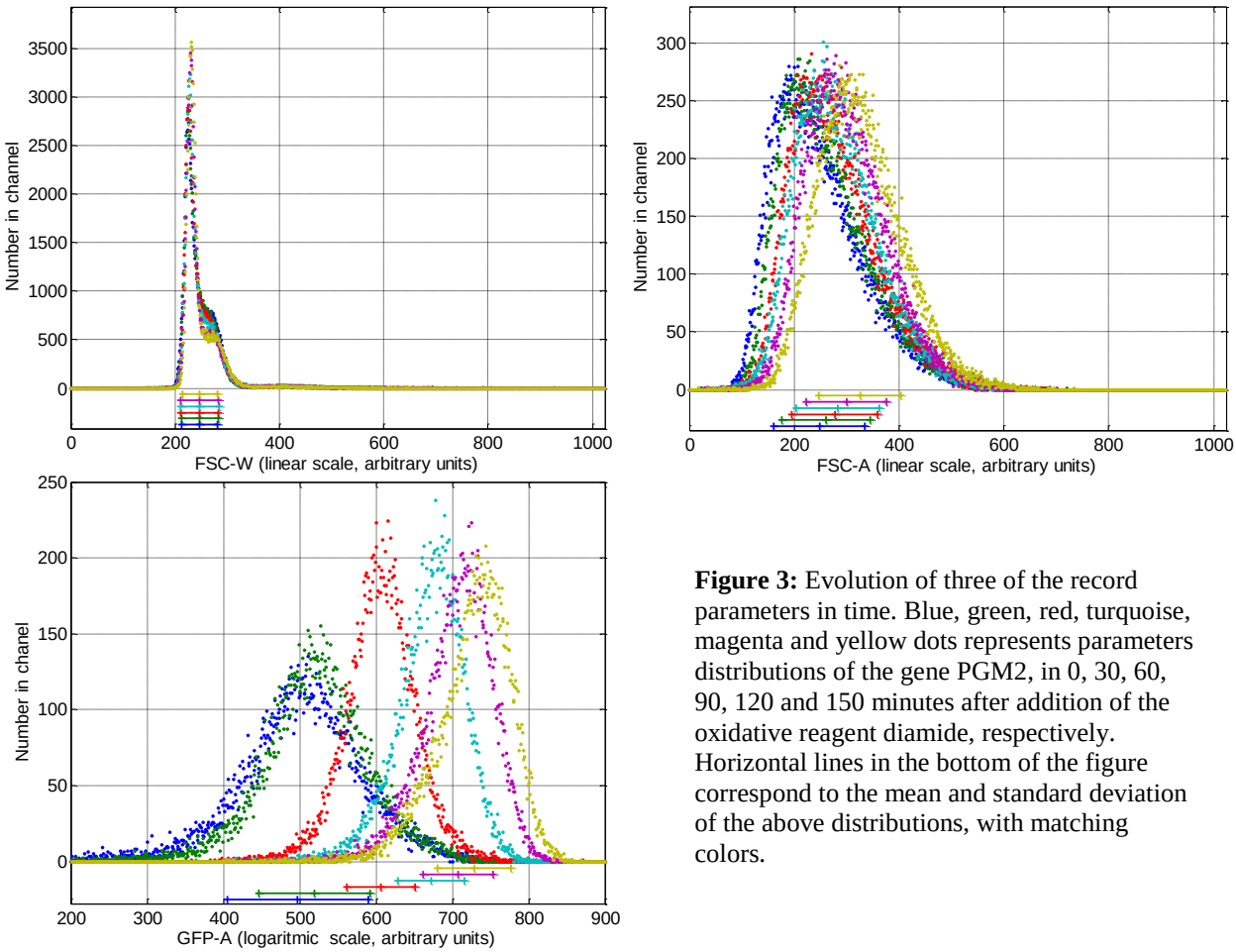

Figure 3: Evolution of three of the record parameters in time. Blue, green, red, turquoise, magenta and yellow dots represents parameters distributions of the gene PGM2, in 0, 30, 60, 90, 120 and 150 minutes after addition of the oxidative reagent diamide, respectively. Horizontal lines in the bottom of the figure correspond to the mean and standard deviation of the above distributions, with matching colors.

Pre-analysis of the flow cytometry results

The raw data recorded by the FACSAria comes from cells with different physiological properties, such as cell sizes or positions in the cell cycle. This can obscure the analysis, as the variability in protein abundance then may reflect the distribution of 'cell types' measured rather than the expression noise in any given type. We therefore select a small but homogenous part of the population by implementing several sequential filters: **1. FSC-W filter**. The usual FSC-A distribution contained two peaks (populations), where the small, high-fluorescence, one is created by cells with buds or by cells in aggregates. The peak value of FSC-W distribution (the left, low-fluorescence, peak) was calculated and the 50% of the cells with the closest FSC-W to that value passed the filter. This filtration eliminated the second peak and therefore discarded most cells with buds and cells in aggregates. **2. FSC-A filter**. The peak value of FSC-A distribution was calculated and the 40% of the cells with the closest FSC-A to that value passed the filter. This filtration created a population of cells that are relatively synchronized with respect to their cell cycle phase, have a comparative cell size and are mostly viable. **3. GFP-A filter**. The 4% cells with the lowest GFP-A values and the 3.5% cells with the highest GFP-A were cut. This filter has two purposes. Some of the GFP-A value recorded are negative, and cutting 4% from the values on the right threw the negative values in all the distributions examined and therefore let us avoid the un-reasonable values. In addition, some of the distributions have distinct outlier, which correspond to dead cells or minor contaminations. Cutting the few percents from both sides eliminates the outlier without changing the main distribution properties. After implementing those three filters we were left with 18,500 recorded cells – fluorescence values.

Normality test of the fluorescence distribution

We used two methods to evaluate the normality of the fluorescence distribution as well as of the distribution of the logarithmically transformed fluorescence values. First we calculated the skewness i.e., the third central moment of the distribution divided by the cube of the standard deviation. Positive and negative skewness indicate long tail to the right and to the left, respectively, while Normal distributions have zero skewness. In addition we created histograms with 1024 bins and used MATLAB's

Curve Fitting Toolbox to fit a Gaussian function to those histograms. Higher order moments, like kurtosis, were not tested.

Calculating the mean, relative mean and standard deviation values

Using the filtered fluorescence distributions we have created three types of data set:

1. Mean values. For each combination of gene, condition and time point we calculated the mean of fluorescence (GFP-A) of the 18,500 cells who past the three filters. Altogether we have (43 genes + ['C'+'CC'] X 2 days) X (11 conditions) X (6 time points per condition) mean values. Only 38 of those 43 genes had sufficiently higher fluorescence, above the auto-fluorescence background. Those genes were taken for the next steps of the analysis. The discarded genes were RPN3, PRS1, RPN7, HSP26 and CYB5. The matrix containing the relative means is shown in figure 5.

2. Relative mean values. We calculated the mean of every distribution and subtracted from each mean value the mean of the control 'C' measured for the same condition and time point. For each gene we subtracted the mean of the control measured in the day it was measured. Next we took the log of the ratio between the mean of each time-point and the first time point, therefore having (43 genes) X (11 conditions) X (5 time points) relative mean values. This process is very similar to the one implemented in micro array analysis.

3. Standard deviation values. We repeated the same procedure as with the mean values, but with the standard deviation of fluorescence instead of the mean, to create (43 genes + ['C'+'CC'] X 2 days) X (11 conditions) X (6 time points per condition) STD values

Noise and noise residuals

We used three definition of noise: $\eta^2 = \sigma^2/\mu^2$, the normalized variance; $\eta = \sigma/\mu$, the coefficient of variation and $\nu = \sigma^2/\mu$, the noise strength, where μ represents the fluorescence mean and σ represents its STD. From those we focused on the first definition. Figure 11 shows that the dependency between σ^2/μ^2 and the mean has three distinct regimes: a $\sigma^2/\mu^2 \sim 1/\mu^2$ regime, a $\sigma^2/\mu^2 \sim 1/\mu$ regime and a $\sigma^2/\mu^2 \sim C$ regime.

In order to find the trend lines we needed to throw the outliers points from the fitting analysis. To exclude these outliers systematically we had preformed an iterative procedure that consists of two steps. At the first step we calculated the best linear fitting using the existing exclusion rule (initialized with no excluded points). This linear fitting was conducted using a fixed slope (-2 or -1, depending on the regime in the log-log graph). At the second step we calculated the vertical distance, of each point, from the trend line and took for our new exclusion rule all points having that distance bigger than 0.5. The process ended upon convergence – no change in the exclusion rule.

To validate the more problematic $\sigma^2/\mu^2 \sim 1/\mu$ dependency we fitted a linear line, across all genes, for each time point of each condition, at the middle regime. In this fitting procedure the slop was not pre-fixed. In all the 11·6 cases the slope, in the log-log graph, was between -0.7 to -1.4, with the average of -1.09. No significant trend was observed across the different time points. Therefore we concluded that the actual slope is indeed -1, with the addition of some experimental noise. The $\sigma^2/\mu^2 \sim 1/\mu^2$ dependency was very clear and hold for each condition and time point.

Every choice of the separation line between the $\sigma^2/\mu^2 \sim 1/\mu^2$ regime and the $\sigma^2/\mu^2 \sim 1/\mu$ affect the fitting curve on both of its sides and therefore also determines the intersection points of those curves with itself. We chose the separation line which makes those intersection points coincide. This border point, log(mean)=6.75, is very close to the median of the log(mean) values of the points for which the background noise is equal or higher than the GFP noise: 6.7742, which strengthen the validity of this border. The separation line between the $\sigma^2/\mu^2 \sim 1/\mu$ regime and the $\sigma^2/\mu^2 \sim C$ was chosen visually.

The vertical distance of the points form the fitted line (each point from the line in the regime it is located in) was defined as the **noise residuals**. We have (43 genes) X (11 conditions) X (6 time points) noise values. The **relative noise residuals** were calculated by subtracting from each noise residual of a specific time point the noise residual of the time point 0. Overall there are (43 genes) X (11 conditions) X (5 time points) relative noise values.

Fluorescence per molecule

The means and standard deviations were measured in units of fluorescence, not in absolute numbers of proteins. Here we assume a proportionality between the two, $\gamma P=F$, where P is the protein abundance, γ is the fluorescence per molecule, and F is total fluorescence. The normalized variance is then independent of the normalizing factor γ because $\sigma_F^2/\mu_F^2=(\gamma\sigma_P)^2/(\gamma\mu_P)^2 = \sigma_P^2/\mu_P^2$, while the means are proportional, $\mu_F = \gamma\mu_P$. In order to estimate the values of γ, we calculated, for each gene, two values. The first was the gene's averaged fluorescence, across the first time points of all the stress conditions, while the second was its protein abundance as measures by Ghaemmaghami et al(29). The mean of the ratio of those values, across all genes, is the estimated $\gamma = 0.16$.

Correlations and dendograms

Dendograms were plotted using the hierarchical clustering algorithm implemented in MATLAB statistical toolbox, using the average linkage option. Distances between entities (genes or time points of conditions) were defined as 1-Pearson correlation across mean, relative mean, noise residuals or relative noise residuals. Dendograms were rendered using the dendogram function in MATLAB.

Expression coherence of motifs

The expression coherence (EC) of a motif is a measurement for the extent to which the co-regulation of genes having that motifs is higher than the co-regulation of those genes with genes that do not have the motif. First, we calculated the Pearson correlation values between genes according to their mean, relative mean or noise residuals or relative noise residuals. We created two value sets. The first contained the correlation values calculated between genes that share the motif and the second contained the correlation values calculated between genes that have the motif and genes that do not have it. The EC was defined as the $\log_{10}$ of the P-value (using a rank-sum test) of the hypothesis that both value sets have the same median.

RESULTS

Shape of fluorescence distributions

Distributions of protein abundance of 43 of S. cerevisiae genes was created using different yeast strains, each having a GFP fused to a 3'UTR of a different protein. From these GFP strains only 38 had sufficient fluorescence intensity above the background. Using flow cytometry, fluorescence of single cells was measured, for each of those strains, in 11 different conditions. From the eleven conditions 8 were stressful while the others were stress relaxation (see Methods). Every gene, at each condition, was measured in 6 time points, during 150 minutes.

The shape of the protein abundance distribution is very informative of the molecular processes that created it. If the final protein abundance was the summation of several random variables (each corresponds to a different up-stream process) we would expect it to have a normal distribution (the central limit theorem). A log-normal distribution will be indicative of a multiplication of several random variables. If one, Gaussian, up-stream process has a significantly higher variance than all the other processes, the resulting distribution will be normal, both if it originated from summation or multiplication of random variables.

The fluorescence distributions possessed both normal and log-normal characteristics. We checked the skewness (defined as the third central moment divided by the cubic standard deviation) of the linear and the logarithmic transformed fluorescence values, in all genes, conditions and time points. Skewness is a measure of the asymmetry of a distribution: skewness<1 is indicative for a long distribution tail to the left, while skewness>1 is characteristic of distributions with a long left tail. Gaussian distribution is characterized in skewness of 0. The average skewness, of linear and log fluorescence values of different genes, across all conditions and time points, are given in figure 4(top, right and left).

There is a clear dependence between the mean abundance of a gene and the average skewness of its linear and log fluorescence values. Low abundant genes tend to have normal distribution – skewness near 0 of their linear fluorescence values and high skewness of their log fluorescence values, indicative of a long tail to the left. High

abundant genes present the opposite trend, having a log-normal distribution – skewness near 0 of their log fluorescence values and low skewness of their linear fluorescence values, which indicate a long tail to the left. Therefore it seems that the higher the protein abundance, the fluorescence distribution becomes more log-normal and less normal. The shift from the normal to the log-normal pattern is continuous: the difference between the logarithmic and the linear skewness is a monotonously decreasing function with the mean abundance, as shown in figure 4(bottom, left). Moreover, genes that are induced in the time course of a condition shift slowly from normal to log-normal distribution (not shown).

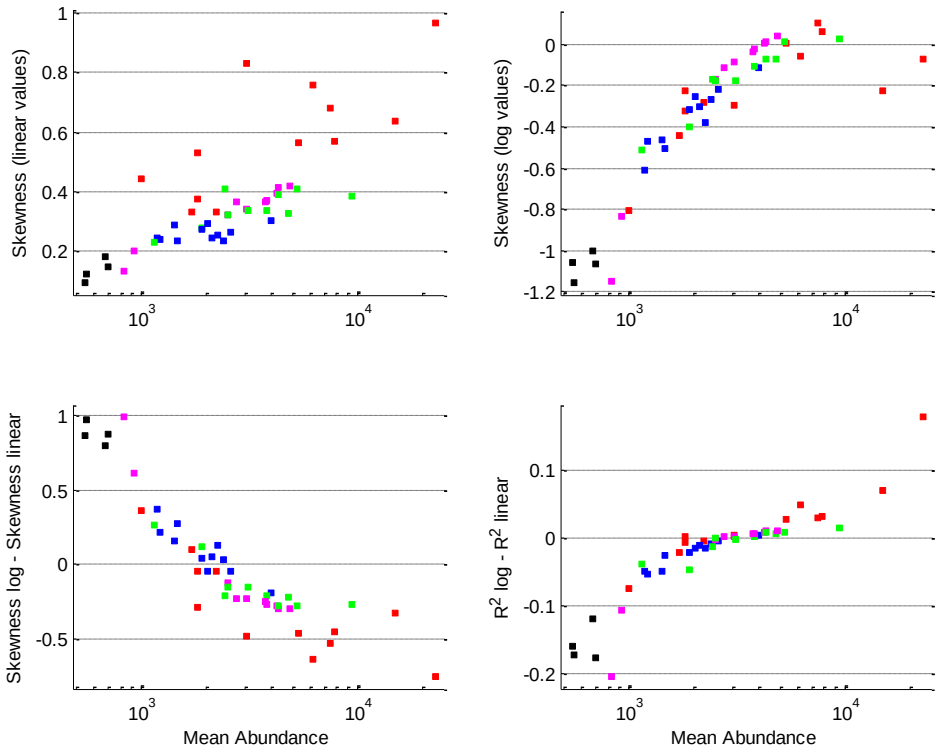

Figure 4. Top left. Mean abundance and skewness of linear fluorescence values, of different genes. **Top right**. Mean abundance and skewness of log fluorescence values, of different genes. **Bottom left**. Mean abundance and difference between skewness of log fluorescence values and skewness of linear fluorescence values. **Bottom right**. Mean abundance and difference between the R^2 of log fluorescence fitting to normal distribution and the R^2 of linear fluorescence fitting to normal distribution. Genes are represented as solid squares and colored according to their module affiliation: red – stress, magenta – proteasome, green – ergosterol, blue – rRNA processing and black – control.

To validate this conclusion we fitted each linear and logarithmic distribution to a normal one, using MATLAB's Curve Fitting Toolbox, and calculated the difference between the R^2 of the logarithmic fitting and the R^2 of the linear fitting, as shown in

figure 4(bottom, right). The results validate our hypothesis: higher abundance indicate better fitting to log-normal than to normal distribution. Interestingly, most stress genes seem to have higher skewness of fluorescence values on the linear scale, as compared to genes from other modules.

Analysis of the mean protein abundance

The mean values of the fluorescence distribution are shown in figure 5.

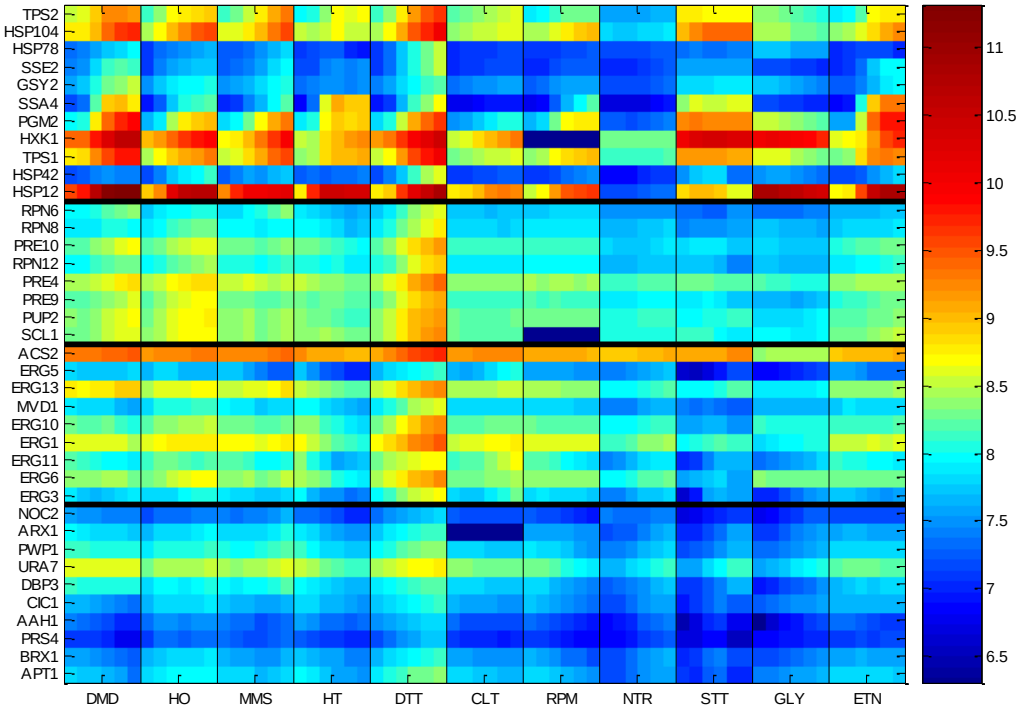

Figure 5: The log of the average fluorescence of the 38 genes analyzed. Each row represents different gene, while the columns correspond to the 11 conditions, each in 6 time points (0, 30, 60, 90, 120 & 150 minutes). Time points of the same condition are ordered consecutively. The genes are clustered in their module affiliation order, from top to bottom: stress, proteasome, ergosterol and rRNA processing. The conditions are given in their abbreviations, as given in the Methods.

We also calculated the relative mean values, corresponding to the log-ratio between the mean in a certain time point and the first time point, after subtracting the background given by the control (see Methods).

Gene clustering and PCA using the mean abundances

The analyzed genes were deliberately chosen from four distinct modules: stress, proteasome, ergosterol and rRNA processing. As the first step we would like to check whether the mean of the protein abundance cluster according to the genes module affiliation. For this purpose we calculated the Pearson correlation of all the genes, across all conditions and time points, using the relative mean values, and preformed average linkage clustering, as shown in figure 6.

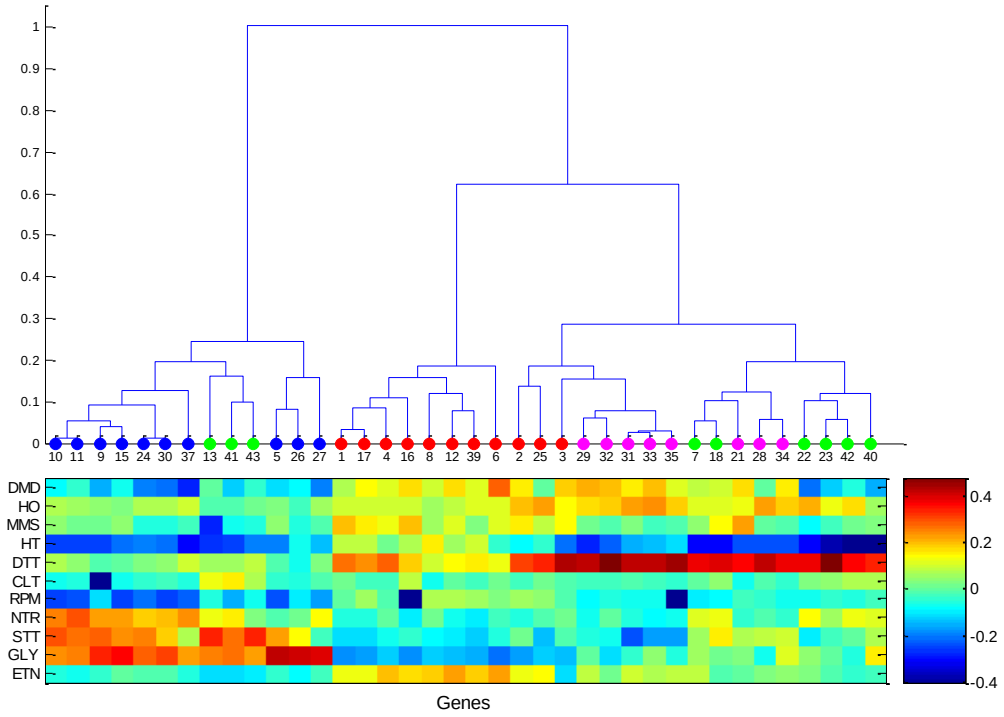

Figure 6: Dendogram of the genes using Pearson Correlation on the relative mean values and clustering by average linkage method. Numbers represent the internal numbering of genes (see Methods). Color of leafs represents the genes module affiliation: red – stress, magenta – proteasome, green – ergosterol and blue – rRNA processing. The bottom panel shows the values of each gene in the last time point of each condition, normalized to all the values of that gene.

As seen, genes largely cluster according to their known module affiliation. The rRNA processing genes are separated from the stress and proteasome ones and the latter two modules are generally located in different clusters. However, there are several expectations. First, several stress genes are clustered with the proteasome genes. Second, the ergosterol genes are divided between the proteasome and the rRNA genes. The second phenomenon is also observed when clustering using the mean values, instead of the relative mean values, or when using single or complete instead

of average linkage. The patterns of the different modules fit the known behaviors of the different modules: The rRNA processing in repressed in stress conditions and induced in stress relaxation conditions while the other module, especially the stress one, display the opposite trend.

Cases in which clustering is inconsist with the module affiliation could result, at least in part, from sensitivity of clustering algorithms to experimental noise. In order to analyze further the effect of module affiliation on the mean protein abundance we preformed a principal component analysis (PCA), which is less sensitive to noisy data. The PCA, which was preformed using the relative mean values, shows a clear separation between the clusters, with some of ergosterol genes in the rRNA proceesing genes era, as seen in figure 7.

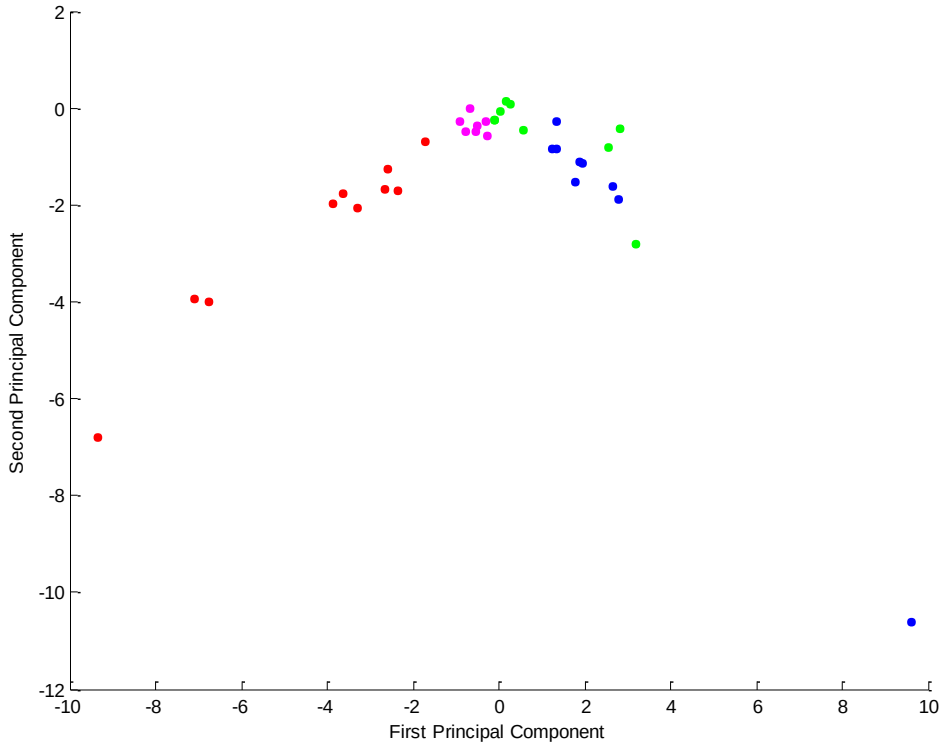

Figure 7: PCA of the genes using the relative mean values. Color coding as in figure 6. The first two principal components capture 90% of the overall variance.

It is interesting to note that the order of the genes on the inverted U shape arc is in correlation with their known response to stress conditions: the stress genes having the largest induction, proteasome genes showing only moderate induction, ergosterol genes moderately repressed and rRNA processing genes being highly repressed.

Comparing these results with the micro array results of Gasch et al., who used similar stressful conditions, reveals that the ergosterol genes tend to cluster just between the proteasome and rRNA processing, as in our results(30). For every pair of genes we could compare their correlation as calculated by the micro array experiment and as calculated by our data set, hence we can compute the correlation between the correlations (CorrBtCorrs). When using the mean values as our data set we get CorBtCorrs=0.51. CorrBtCorrs jumps to 0.65 if our correlation is calculated using the relative mean values (which correspond to the fact the mRNA values are also relative).

Another measurement of the effect of module affiliation on is DiffCorrs: the difference between the mean correlation between genes from the same module and the mean correlation between genes from different modules. For the mRNA data we get DiffCorrs=0.82 (using the same 38 genes) and using the relative mean values of our data set we get DiffCorrs=0.58. Those are indeed comparable values, though our data suggest less module affiliation effect.

Noise analysis

We measured noise as the normalized variance, σ^2/μ^2, where μ and σ are the mean and STD of the fluorescence distribution. The noise levels of all the genes in all the conditions and time points are given in figure 8.

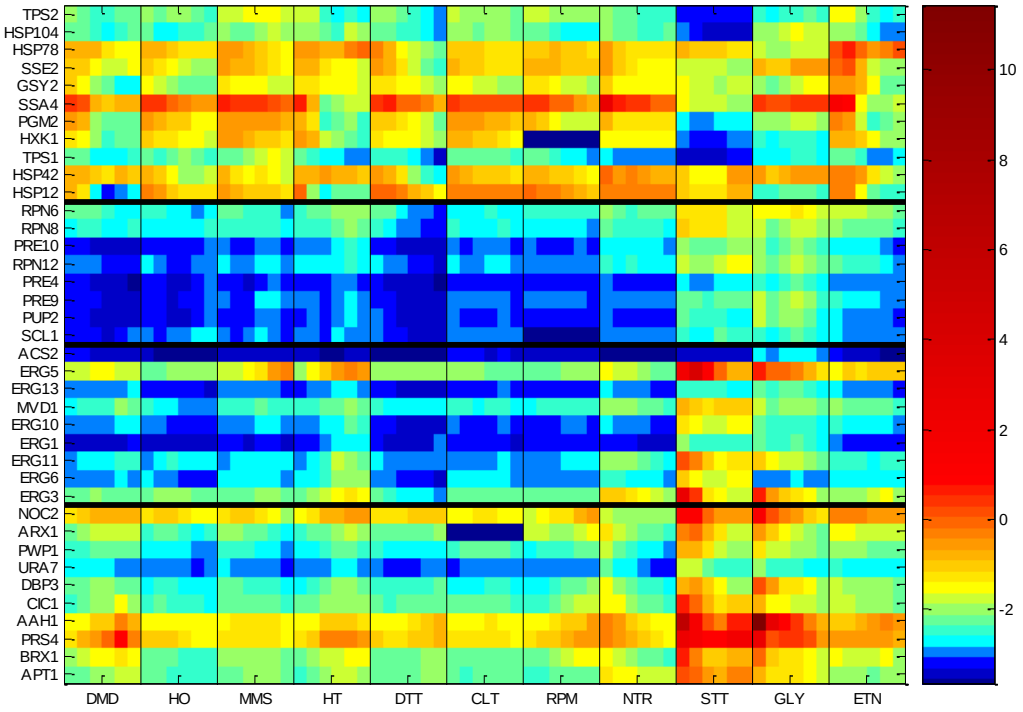

Figure 8: The log of the noise level of the 38 genes analyzed. Each row represents different gene, while the columns correspond to the 11 conditions, each in 6 time points (0, 30, 60, 90, 120 & 150 minutes). Time points of the same condition are ordered consecutively. The genes are clustered in their module affiliation order, from top to bottom: stress, proteasome, ergosterol and rRNA processing. The conditions are given in their abbreviations, as given in the Methods.

Analytical perspective

To try and understand the mechanisms underlying the relationships between noise strength and protein abundance, we analyzed theoretically the expected behavior for several stochastic processes which control protein abundance. For simplicity, we consider each noise source separately, focusing first on stochastic processes effecting protein expression and second on stochastic protein partition during cell division. Notably, processes which are independent contribute to the noise strength in an additive manner(7). Although this assumption does not always hold (e.g. global noise tends to influence several upstream factors, hence creating some dependency between them, see Pedraza et al(24)), such correlation will typically have only second-order

effect on noise strength. Finally, following previous studies, we assume that the different processes involved in transcription and translation are poissonian.

Stochasticity in protein expression

We tried to understand different noise patterns using the mathematical framework of Johan Paulsson(7) (see Introduction). The mean and normalized variance of a certain protein P, which has a dominant extrinsic noise source S, will follow: $\mu(P) \sim f_{S,P}(\langle S \rangle)$, $\sigma(P)^2/\mu(P)^2 \sim a_{S,P}/\langle S \rangle$. We will try to explain how this simple model can create different dependencies between the normalized variance and the mean.

A $\sigma^2/\mu^2 \sim C$ dependency: Let us suppose that for all proteins the dominant source of noise is a global entity, such as ribosomes, polymerases or proteasomes abundance. The mean of the protein abundance will not be equal for different proteins, because of the distinct $f_{S,P}$ function attributed to each of them. However, there is no reason to assume any dependency between the coefficients $a_{S,P}$ the mean abundance, and hence $\sigma^2/\mu^2 \sim C$ (figure 9a). Moreover, if the prime source of noise would be the fluctuation in the abundance of regulators, such as transcriptions factors, common to all genes in a certain module, that regulator should be treated as a global source – all genes belong to that module will have a σ^2/μ^2 which is not dependent on μ, when measured under the same condition (figure 9b).

A $\sigma^2/\mu^2 \sim 1/\mu$ dependency: There are several stochasticity mechanisms that can create this dependency. The trivial one is the intrinsic noise, originating form the Poission process, describing the production and elimination of the protein (figure 9c). Second is a mechanism opposite for the one described for creating the $\sigma^2/\mu^2 \sim C$ dependency. Assume that main extrinsic noise contributor is of the same type for all genes such that each gene has different mean abundance of that entity, but roughly the same proportionality coefficients $a_{S,P}$. That is, the mean protein abundance is proportional to the mean abundance of that entity, across all proteins: $\mu(P) \sim C \cdot \langle S \rangle$, where C is a constant, or at least random variable with relatively small variance. A good example to such scenario will be the mRNA abundance – genes produce diverse abundance of mRNA but, apart from less common post transcriptional regulation, there are

comparable proportionalities between the mRNA levels and the protein levels. In this case the normalized variance follows: $\sigma(P)^2/\mu(P)^2 \sim a_S/\langle S\rangle = a_S \cdot C/\mu(P)$ (figure 9d).

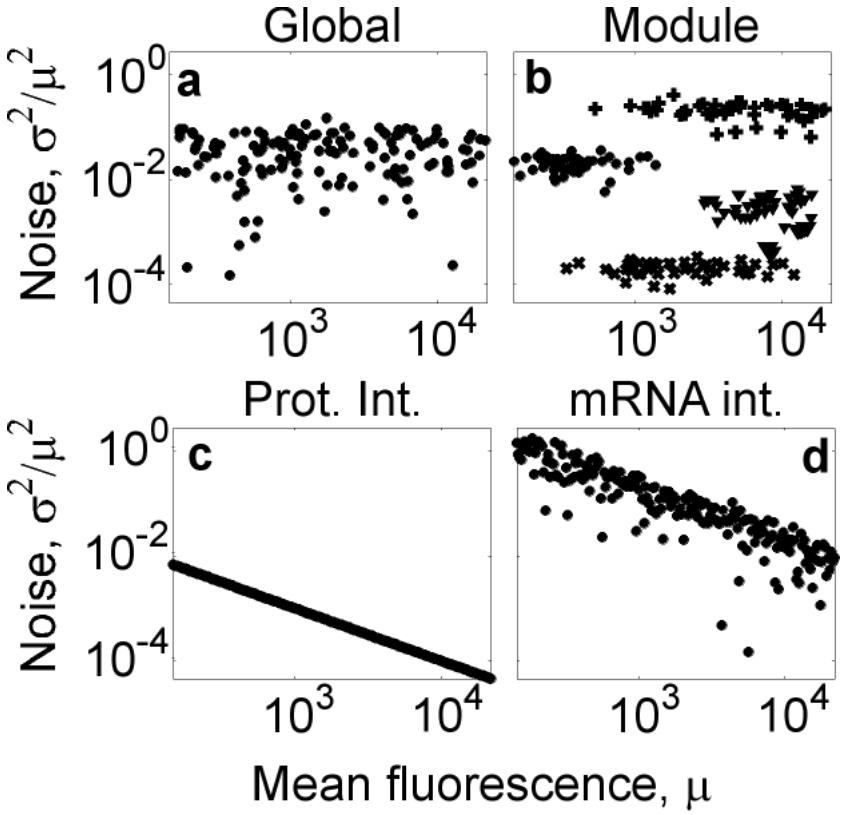

Figure 9. The expected dependencies between the noise and the mean of protein abundance (fluorescence), considering different stochasticity sources. All graphs were simulated using MATLAB. In **a,c-d** we show 200 genes with various protein abundances – selected randomly on the logarithmic x-axis. In **c** we show 4 modules, each containing 50 genes, whose mean abundances are distributed normally (on the logarithmic x-axis) around a certain center. The noise level was calculated using the formulas given below, where the parameter 'c' was randomly chosen, for each gene, from the distributions given below. **a**. Global extrinsic noise of genes with various protein abundances, obeying $\sigma^2/\mu^2 \sim c/\langle G\rangle$, where $c \sim N(1,1)$ and $\langle G\rangle=30$. **b**. nodule extrinsic noise of genes belonging to different modules, obeying $\sigma^2/\mu^2 \sim c/\langle R\rangle$, where $c \sim N(1,1/3)$ and $\langle R\rangle$ is equal to 5 ('X' signs), 50 (triangles), 300 (circles), and 5000 ('+' signs). **c**. Protein intrinsic noise of genes with various protein abundances, obeying $\sigma^2/\mu^2 \sim 1/\mu$. **d**. mRNA intrinsic noise of genes with various protein abundances, obeying $\sigma^2/\mu^2 \sim c/\mu$, where $c \sim N(190,95)$.

The combination of two sources can create more complex trends. If the two major origins of noise are the mRNA level and a global factor, such as the ribosome or the polymerases abundance, the mean and normalized variance will follow: $\mu(P) \sim C \cdot \langle M(P)\rangle \cdot \langle G\rangle$ or $\mu(P) \sim C \cdot \langle M(P)\rangle$, $\sigma(P)^2/\mu(P)^2 \sim a_{M(P)}/\langle M(P)\rangle + a_{G,P}/\langle G\rangle$, where M and G represent the mRNA and the global factor abundances, respectively and C represent the fixed (or low variance) proportionality coefficient. $\langle G\rangle$ is fixed for all genes, while $\langle M(P)\rangle$ is generally different. In proteins, which have low mRNA abundance (low abundant proteins) the dominant noise factor will be the mRNA

abundance and hence the plot will have a dependency of $\sigma^2/\mu^2 \sim 1/\mu$. Proteins which have high mRNA abundance (high abundant) will have the global factor (ribosomes or polymerases abundance) as the main source of noise and the dependency will become $\sigma^2/\mu^2 \sim C$.

A $\sigma^2/\mu^2 \sim 1/\mu^2$ dependency: Having a constant standard deviation for all the genes, regardless their mean will create this dependency.

Stochasticity in protein partition during cell division

The process of cell division, followed by protein partition between the daughter cells, can create a significant noise level. There are two types of noise it can produce: a binomial partition noise and noise originate from un-equal division.

First we will assume a perfect division to equal size daughter cells. In this case there will be only a binomial partition noise. The protein abundance in each of those cells will follow a binomial distribution, with partition coefficient p=0.5; hence $\mu=np=0.5n$, $\sigma^2/\mu^2=np(1-p)/n^2p^2=(p(1-p)/p^2)/n=1/n=1/2\mu$, therefore creating a dependency of $\sigma^2/\mu^2 \sim 1/\mu$.

However, several cells, as the budding yeast, does not split evenly – the daughter cell is smaller the mother one. Therefore, the partition coefficient p will be smaller than 0.5 (or higher, depending on the point of view). In this case their will be an additional noise contribution, equal to $4\cdot(0.5-p)^2$. The extrinsic noise will be even higher if we assume stochasticity in the division process itself, which will result in p as a random variable. The most simple way to take this effect into account, is by assuming $p \sim N(\mu_{prob},\sigma_{prob})$, where partition coefficient mean, μ_{prob}, can range from 0 to 0.5 and the partition coefficient STD, σ_{prob}, is determined by the extent of fluctuations in the cell division. Higher σ_{prob} will create higher the extrinsic noise. The un-equal division noise as a function of μ_{prob} and σ_{prob}, calculated using stimulations of cell division and proteins partition between daughter cells, is giving is figure 10(left). This noise, behaving different from the binomial partition one, does not depend on the mean protein abundance. As a result the total noise from protein partition will behave as

$\sigma^2/\mu^2 \sim 1/\mu$ at low protein abundance and as $\sigma^2/\mu^2 \sim C$ at high protein abundance, as shown in figure 10(right).

There is relationship between the two type of protein partition noise and the intrinsic and extrinsic sources of noise, as defined for the two reporter experiments. The binomial partition noise will be measured as intrinsic noise, while the un-equal division noise will be measured as extrinsic one.

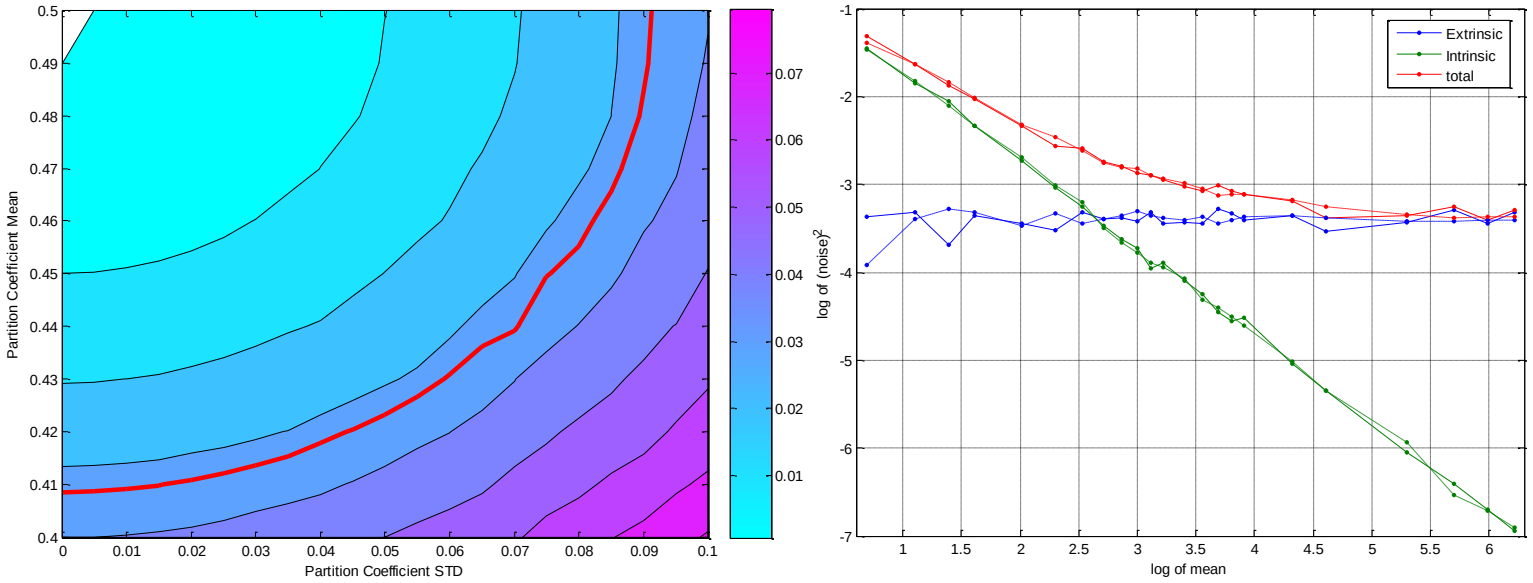

Figure 10. Left. The un-equal division noise as function of μ_{prob} and σ_{prob}. For each combination of μ_{prob} and σ_{prob}, the partitions of various amounts of protein were simulated for 2500 cells. The intrinsic and extrinsic contributions to the overall noise were calculated using the formula described at Raser et al(23). The un-equal division noise (extrinsic noise) did not depend on the total amount of proteins. Red curve corresponds to extrinsic noise of 0.0334, which was the lower bound measured experimentally in the $\sigma^2/\mu^2 \sim C$ dependency regime, as explained in the text below. **Right**. Binomial (Intrinsic), un-equal division (extrinsic) and total noise simulated for μ_{prob} and σ_{prob} values located at both ends of the red curve (solid and dashed lines). Notice that the noise patterns are identical – not only the extrinsic contribution, but also the intrinsic one. In all cases the partition of n (varying on the X-axis) proteins was simulated for 2500 cells.

Noise patterns in the experimental data

The dependencies between of the normalized variance and the noise strength on the mean abundance, for all genes in all conditions and time points, are given in figure 11.

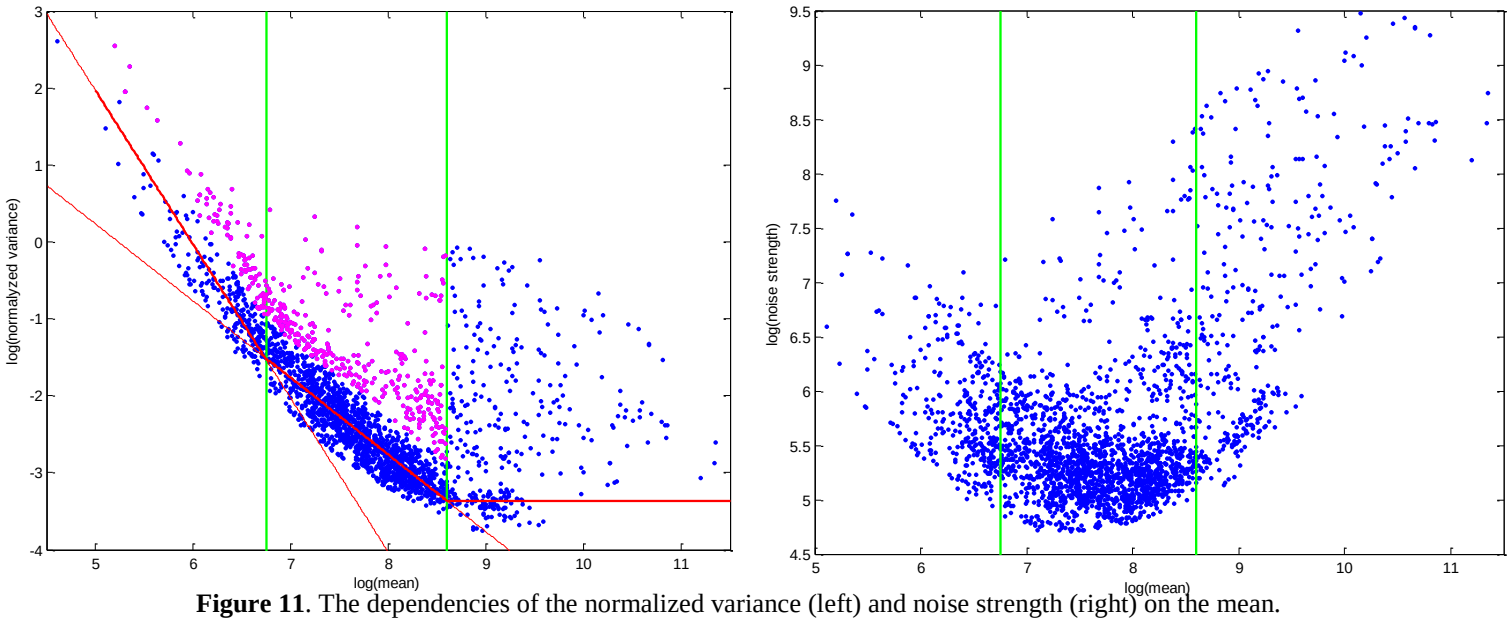

Figure 11. The dependencies of the normalized variance (left) and noise strength (right) on the mean. Green line separates the three regimes of dependencies, as explained in the text. Red lines are the fitted trends in each regime. Points colored with magenta are ones who were excluded from the fitting, see Methods.

There are three distinct regimes of dependencies between the noise and the mean. The decision of regimes border was made as described in the Methods. In each regime the noise is dominated by different factor and therefore its dependency on the mean is different, as given in table 3.

Regime	Dependencies	Slop in the normalized variance log-log graph	Slop in the noise strength log-log graph
Right	$\sigma^2/\mu^2 \sim 1/\mu^2$	-2	-1
Middle	$\sigma^2/\mu^2 \sim 1/\mu$	-1	0
Left	$\sigma^2/\mu^2 \sim C$	0	1

Table 3. The different dependencies and slopes, existing in the distinct regimes.

The total measured noise can be decomposed into the following terms:

$$\frac{\sigma_P^2}{\mu_P^2} = \underbrace{\frac{\sigma_{background}^2}{\mu_P^2}}_{\text{Background noise}} + \underbrace{\frac{1}{\mu_P H_{P,P}}}_{\text{Protein Intrinsic noise}} + \underbrace{\left(\frac{1}{\mu_M H_{M,M}} + \eta_{up}^2\right)}_{\text{mRNA intrinsic noise}} \underbrace{\frac{H_{P,M}^2}{H_{P,P}^2} \frac{H_{P,P}/\tau_P}{H_{M,M}/\tau_M + H_{P,P}/\tau_P}}_{\text{mRNA upstream noise}} + \underbrace{\eta_{down}^2}_{\text{mRNA downstream noise}}$$

where $\sigma_{background}$ is the standard deviation of the constant background fluorescence, η^2_{up} represent the extrinsic noise from factors up-stream to the mRNA abundance and

η^2_{down} corresponds to extrinsic noise from factors down-stream to the mRNA abundance. All other signs are as in Paulsson's basic equation, where P and M represent protein and mRNA, respectively. We will assume the biologically reasonable $H_{P,P}^2=H_{P,M}^2=H_{M,M}^2=1$.

The measured mean and standard deviation was that of the fluorescence and not of the protein abundance. We compared the published protein abundance data set(29) to our fluorescence measurements and found out that the fluorescence of a single GFP is approximately 0.16 arbitrary units (see Methods).

The $\sigma^2/\mu^2 \sim 1/\mu^2$ dependency: The fluorescence width of low abundant proteins, populating the right regime, is highly affected by the background fluorescence. Therefore, the first term in the equation above is the dominant one in that regime, creating the $\sigma^2/\mu^2 \sim 1/\mu^2$ dependency. The intersection point of the fitted line, in that regime, with the Y-axis should correspond to the log value of $\sigma_{background}^2$. The average of the background fluorescence was 131,530, which corresponds to expected intersection point of 11.79, while the actual intersection was 11.97. Indeed, a very good match. This experimental noise source decay rapidly with the mean abundance and becomes negligible for protein with higher abundance.

The $\sigma^2/\mu^2 \sim 1/\mu$ dependency: The middle trend cannot originate from global, module or un-equal division sources because those sources create a dependency of $\sigma^2/\mu^2 \sim C$. Moreover, neither protein intrinsic noise, nor binomial partition noise, can explain the observed results, although both create a $\sigma^2/\mu^2 \sim 1/\mu$ dependency: The intersection of the fitted curve, in this regime, with the Y-axis is 5.237. Therefore, the coefficient of the 1/<fluorescence> term is exp(5.237)=188. Converting this value to protein number, the coefficient of 1/<protein> is 188/0.16=1175. If protein intrinsic noise were the primary noise source we would expect a coefficient of ~1 by the Possion process model. The binomial partition model will give such high coefficient only if p would be equal to 1/1176, which is unreasonable.

The only plausible origin of the $\sigma^2/\mu^2 \sim 1/\mu$ dependency is the effect of mRNA abundance, as explained above. In order to validate this hypothesis we calculated the expected coefficient and compare it to the observed one. The expected coefficient is

the multiplication of the proportionality between mRNA and protein abundance and the time averaging factor. We estimated that proportionality by using the protein and mRNA abundances published by Ghaemmaghami et al.(29) end Greenbaum et al(31), respectively. In order to calculate the time averaging factor, we used the mRNA decay rate data set, published by Wang et al.(32) and assumed that protein life-time is dominated by the cell-cycle period; hence we estimated it as 100 minutes. The expected coefficient (averaged for all genes) is:
$(\mu_{protein}/\mu_{mRNA})\cdot(1/\tau_{protein})/(1/\tau_{mRNA}+1/\tau_{protein})=1105$,
which is very close to the measured coefficient – 1175.

However, the hypothesis that mRNA intrinsic noise is the origin of the $\sigma^2/\mu^2\sim1/\mu$ dependency does not fit the experimental finding of previous works, which demonstrated that the processes down-stream to the promoter activation (that is, transcription process, mRNA degradation, translation process and protein degradation) contributes only few percent up to less than 50% of the overall noise(17, 19, 20, 23, 24). Moreover, Raser et al. demonstrated that the correlation between the abundance of proteins, express from different promoters, is considerably high, indicating that a global source of noise has a dominant role in determining the noise level(23).

<u>The $\sigma^2/\mu^2\sim C$ dependency:</u> The left trend can originate from three different sources: global, module and un-equal division noise, all creating the $\sigma^2/\mu^2\sim C$ dependency. If module affiliation was playing a major role in determining noise level we would expect proteins belonging to the same module to display no dependency between σ^2/μ^2 and μ. As shown in figure 12 this does not hold – genes within a module share the $\sigma^2/\mu^2\sim1/\mu$ and the $\sigma^2/\mu^2\sim1/\mu^2$ dependencies. Hence we can conclude that module affiliation has only a secondary effect on the noise level of a protein and is not responsible for the observed trend.

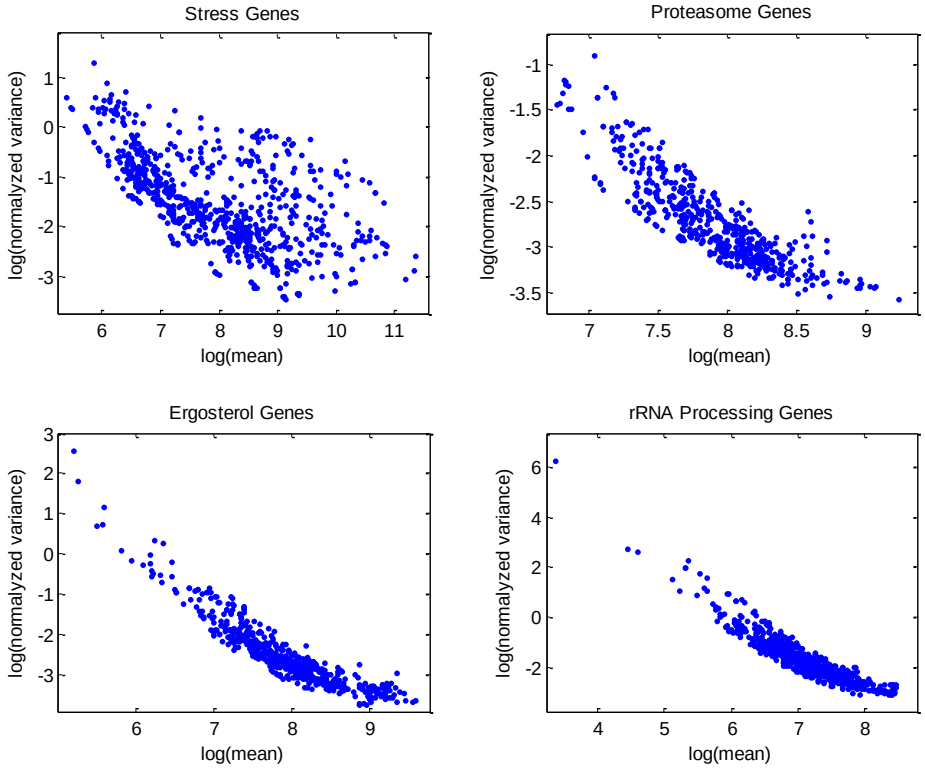

Figure 12. Proteins belonging to the same module do not have a σ^2/μ^2~C dependency, as expected if module affiliation was to play a major role as noise source.

The existence of dominant global factor, such as polymerases, ribosomes and proteasomes can account for this dependency. The global factor can affect the protein abundance directly, down-stream or parallel to the mRNA abundance, or indirectly through the mRNA contribution, being up-stream element in its cascade. In the latter case, the averaging factor attributed to the mRNA source of noise will dump also the up-stream factors. The σ^2/μ^2 lower bound in that regime is 0.0334. That value corresponds to the stochasticity expected by having several dozens copies of the 'mRNA down-stream' global factor or only several copies of 'mRNA up-stream' global factor (about 29 and 4, respectively, if neglecting the susceptibility and time averaging of that global factor). RNA polymerases, being up-stream to the mRNA abundance, is not a good candidate because cells have more than several of those complexes. Ribosomes and proteasome complexes are located down-stream to the effect of the mRNA. However, ribosomes are clearly present in more than several dozens of copies. Because we analyzed relatively small number of genes, and there are only few genes at the actual lower bound of the σ^2/μ^2~C regime, it could be that the bound is determined by a low abundant transcription factor that regulates those

genes. This hypothetical transcription factor, being an up-stream element to mRNA abundance, should be present in no more than several copies.

Another explanation, for the $\sigma 2/\mu 2 \sim C$ dependency, could be the extrinsic noise originates in cell division, as described in the partition model above. The solid curve in Figure 10 (right) represent combination of μprob and σprob values, which create an extrinsic noise of exp(-3.4)=0.0334, as the measured lower bound. The expected values of μprob (0.4 to 0.5) and σprob (0 to 0.1) are the biologically reasonable. Obviously, the contribution of low abundant global entity and un-equal cell division is additive and the observed lower bound can be a function of both.

The noise residuals and their dependency on module affiliation

We hypothesized that the secondary effect of module affiliation may be hidden in the difference between the actual σ^2/μ^2 of certain protein abundance and the value predicted by the trend line, given the mean protein abundance. We term those differences noise residuals. The noise residuals were calculated using the log values of both the mean and the normalized variance. Subtracting from each such noise residual the noise residual of the same protein, in the same condition, at time point 0 gave the relative noise residuals.

In order to check whether module effects are indeed hidden at the residuals values we averaged, for each gene, the noise residuals it exhibited in each condition and time point. The results are given in figure 13.

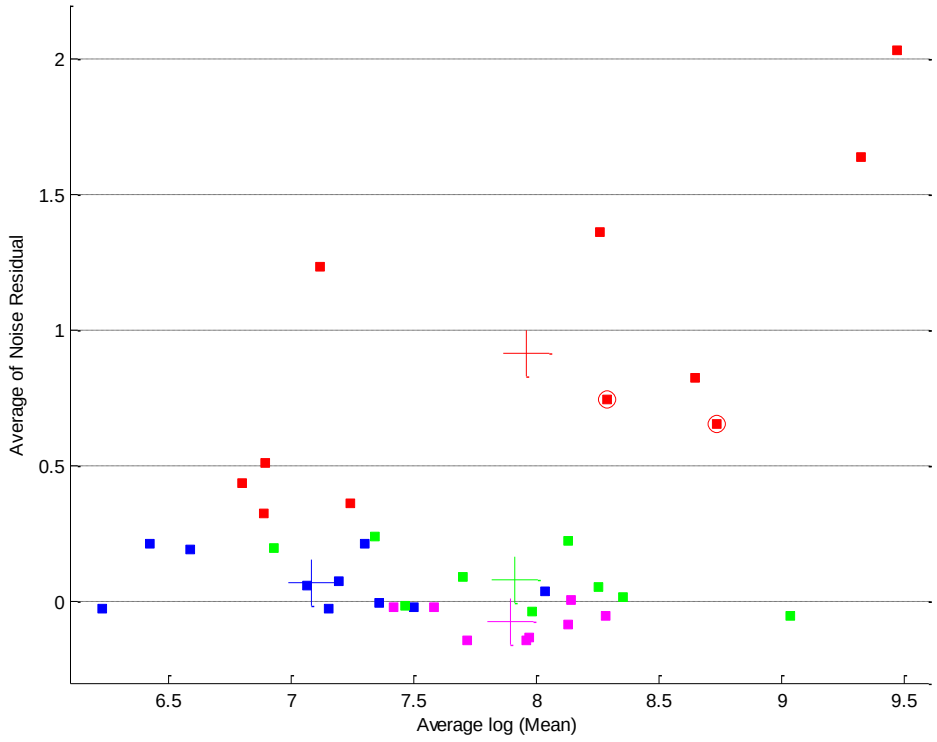

Figure 13: The gene average of mean and noise residuals. Genes are represented as solid squares and colored according to their module affiliation: red – stress, magenta – proteasome, green – ergosterol and blue – rRNA processing. Circled squares are the genes TPS1 and TPS2, see Discussion. The '+' signs represent the averages of a module – averaging on all the conditions, time points and genes belonging to the same module.

The stress genes have high noise residuals, while the proteasome genes are less noisy than the other, although both have relatively the same average of means. Both trends

are highly significance (see table 4), even if controlling for multiple hypothesis using the Bonferroni method. The ergosterol and rRNA processing genes have the same range of noise residuals, despite having very different average of means.

	Stress	Proteasome	Ergosterol	rRNA Processing
Stress		10^{-140}	10^{-118}	10^{-119}
Proteasome			10^{-24}	10^{-16}
Ergosterol				0.34

Table 4. P values, obtained by Rank-Sum tests, of the hypothesis that the median of the noise residuals of two modules is equal.

These results demonstrate the significance of module affiliation in determining the noise residuals. We also tested the correlation of the genes average of noise residuals and other genetic and genomic properties:

- Chromatin properties such as nucleosome occupancy, acetylation and methylation, taken from pokholok et al.(33) and from sequence-based prediction of nucleosome occupancy (Eran Segal, personal communication).
- Chromosomal location – chromosome affiliation and distance from the centromers and the telomers.
- Connectivity in the protein-protein interaction graph, taken from von Mering et al.(34)
- Predicted expression level based on tRNA adaptation index – in S. cerevisia and across eight of the ascomycotic species taken from Orna Man (personal communication) and from dos reis et al.(35)
- Number of promoter motifs regulating each gene – experimental, taken from Harbison et al. and computational, taken from Kafri et al.(36) and from Michal Lapidot (personal communication)
- Phylogenetic distance of each gene from *S. cerevisiae* to the last common ancestor with *Candida glabrata*, taken from Amir Mitchell (personal communication)
- Expression divergence for each genes across several yeast species, taken from Itai Tirosh (personal communication)
- Presence of TATA motif in the promoter, taken from Basehoar et al.(37)
- Fitness of the gene deleted strain, taken from Steinmetz et al.(38)

- mRNA decay rate, taken from Wang et al.(32)

In all those case no correlation could be established. This may be attributed to the relatively small number of genes analyzed. Examining correlation is even more problematic in our data set because of the large effect of module affiliation on noise residuals – the correlations should be checked separately in each module to avoid module affiliation acting as a confounding variable in a specific correlation found. One example to the above issue is the correlation found between the existence of TATA sequence in a promoter of a gene and its high noise residual. It turns out that all, but one, of our stress genes contain that sequence while only 6 genes from other modules include it, therefore creating a biased and misleading correlation. Another example is the finding that dispensable genes have much higher noise residuals than essential ones. However, this trend originates from the fact the all our chosen stress genes were dispensable, while all, but one, of our chosen proteasome genes were essential.

Gene clustering and PCA using the noise residuals

An alternative way of measuring the effect of module affiliation on the noise residual is by checking whether the genes' clustering using those values is consistent with the module affiliation. First, we preformed a PCA on the relative noise residuals, as shown in figure 14.

Module affiliation seems to have a significant effect on the values of the second principal component. Modules are ordered linearly, on this component axis, according to induction by the stress conditions – increasing values correspond to increased induction. However, the noise residuals seem to have a non-module-related component(s) which is responsible for the overlap between the regimes of each module. The module mixing effect of this component is visual when plotting the hierarchical dendogram of the genes, using the average linkage method, as shown in figure 15. The basic module affiliation is clearly visible, but the clustering reveals large distortions from the module-related structure.

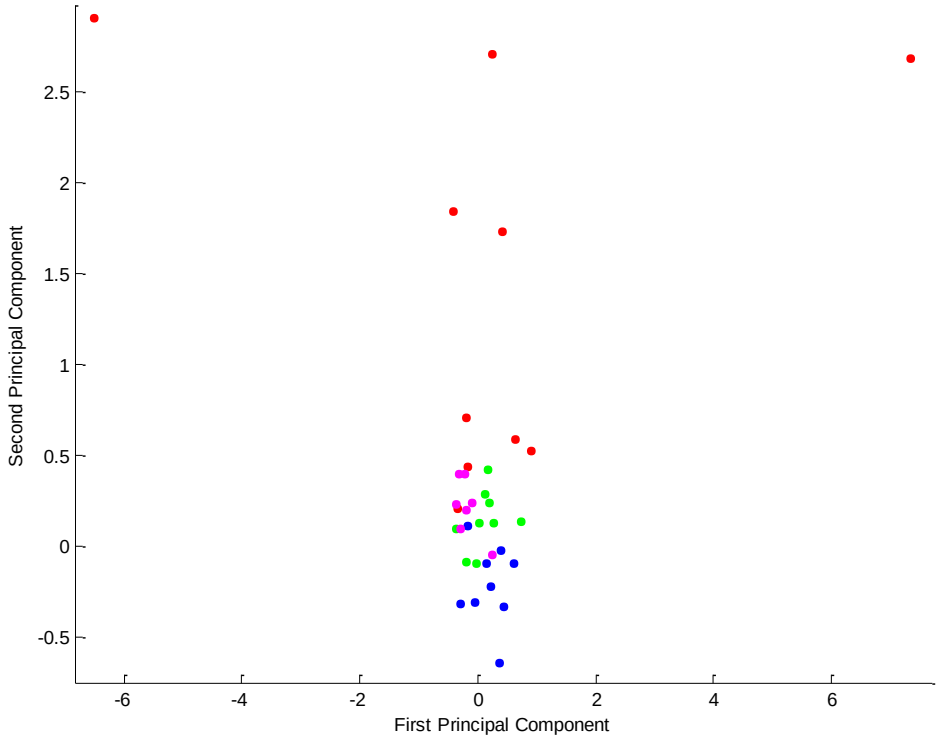

Figure 14. PCA of the genes using the relative noise residuals values. Color coding as in figure 13. The first two principal components capture 79% of the overall variance.

The lesser effect of module affiliation on noise residuals, as compare to its effect on the mean abundance, can be seen in another two factors. Using Gasch et al. data set(30) and the relative noise residuals we get a non-significant CorrBtCorrs (see above) of 0.13 (or 0.14 when using the noise residuals). CorrBtCorrs=0.2 when using the relative mean values and the relative noise residuals values.

In addition, there is a small DiffCorrs (see above) when using the relative noise residuals: 0.16. This DiffCorrs value is much smaller than the ones found using Gasch et al. mRNA data or the relative protein abundance. However, randomizing the module affiliation labels for the genes, for 1,000,000 times, and calculating DiffCorrs did not give a value higher than 0.15, indicating that after all the module affiliation has a dominants contribution to the noise residuals.

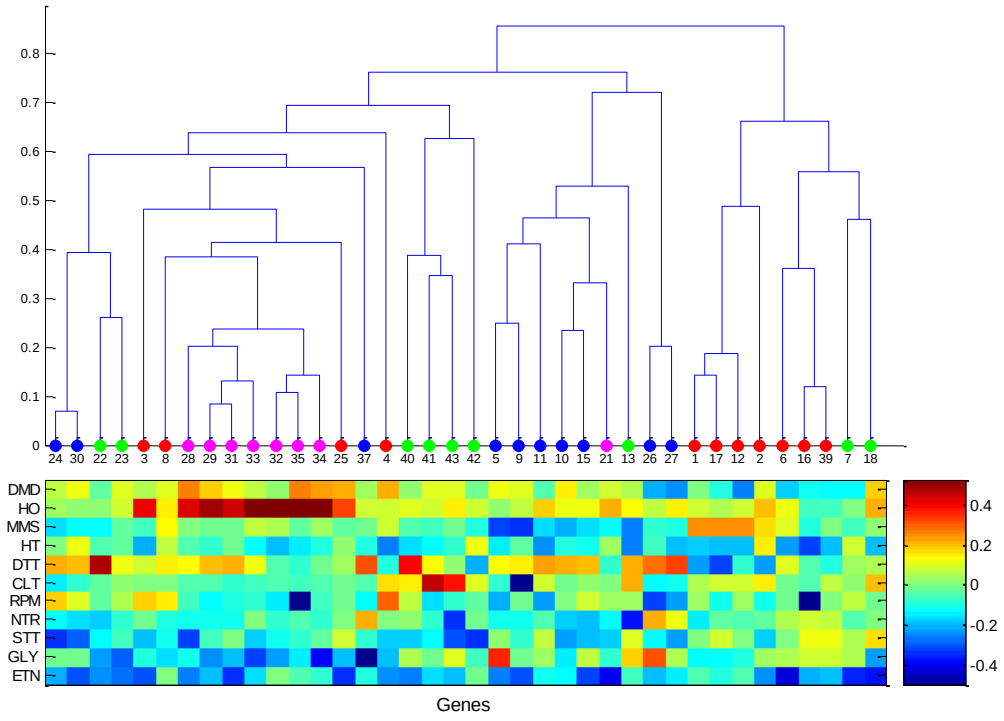

Figure 15: Dendogram of the genes using Pearson Correlation on the relative noise residuals values and clustering by average linkage method. Numbers represent the internal numbering of genes (see Methods Color coding as in figure 13. The bottom panel shows the values of each gene in the last time point of each condition, normalized to all the values of that gene.

Noise change across time points

After establishing the dominant role of module affiliation in determining noise residuals we can ask whether the noise residuals of different modules behave differently in time, when they are perturbed with different conditions. The changes in mean and residuals, across time, for all modules and conditions are given in figure 16, after normalizing to the first time point.

The pattern of mean and noise residuals in time is very diverse. For some conditions, such as heat-shock, nitrogen depletion, glycerol and ethanol all modules seem to share the same general noise residual pattern, although having significantly different mean values pattern. Other conditions, as diamide, MMS and clotrimazole, create different, or even opposite, noise residuals patterns, for different modules. The variety of behaviors in time presented by these conditions emphasizes the fact that the response of the mean of the protein abundance itself is not sufficiently describing the full effect of a perturbation – the behavior of noise in time provides independent information on

each condition. In the nitrogen depletion and stationary condition the noise residuals, of all modules, seem not to change significantly.

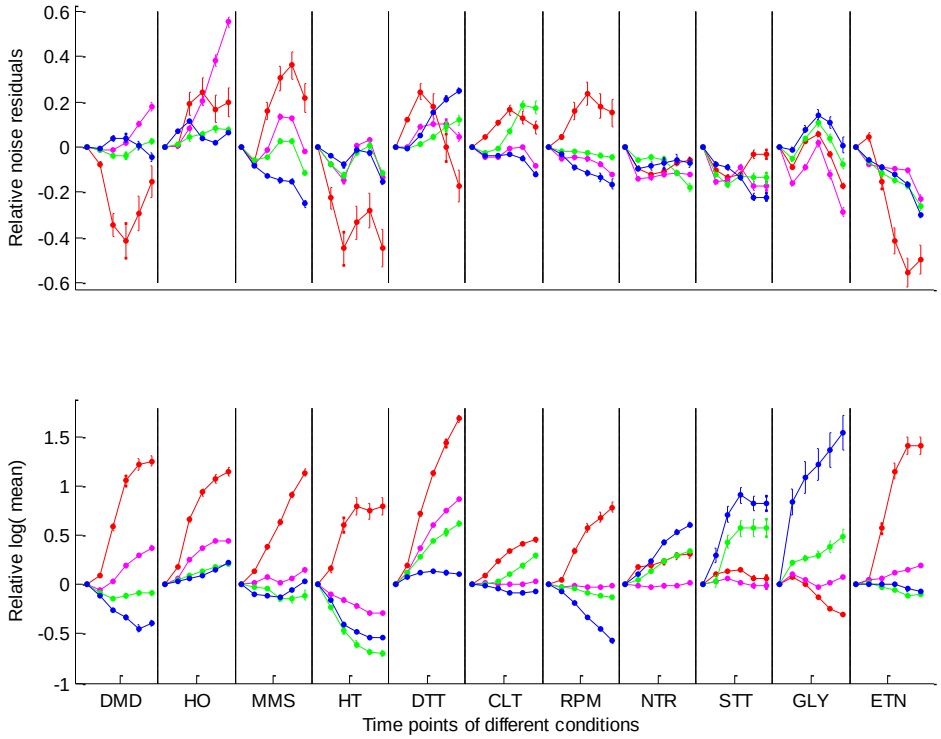

Figure 16: The time points' average of the normalized mean and noise residuals, given in error-bars. Modules are colored as in figure 13.

The overall patterns of the means fit the well known behaviors of the different modules. In the stress conditions the stress module is highly induced, the proteasome is mildly induced, the ergosterol in mildly repressed and the rRNA processing is highly repressed. In the stress relaxation conditions the modules display the opposite patterns.

Conditions clustering using the mean values and the noise residuals

Clustering the conditions, as we did with the genes, can be very informative for the understanding of their influence on the cell. Using the relative mean values we created the dendogram shown in figure 17.

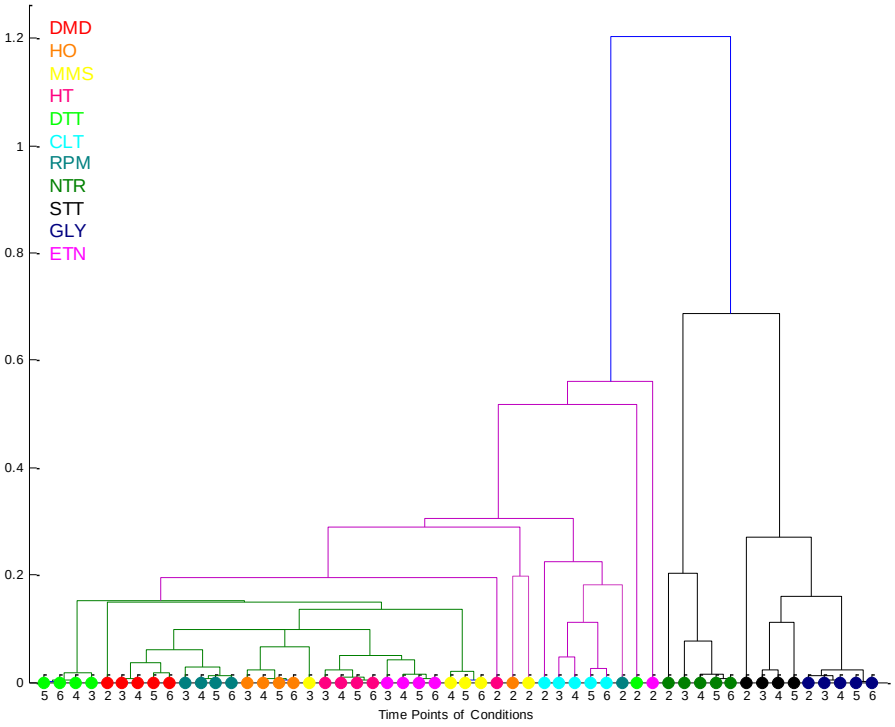

Figure 17: Dendogram of the time points and conditions using Pearson Correlation on the relative mean values and clustering by average linkage method. Numbers represent the time point – starting from 2 because all values are compared to the first time point (see Methods). Leafs color corresponds to the condition as written, in abbreviations (see Methods), in the upper left part of the graph. The lines color corresponds to the three clustered seen, see detailed in text.

As seen, all the stress relaxation conditions are clustered together (black cluster), and sub-cluster to the specific conditions. The pink cluster corresponds to all the second time points of the stress conditions together with the Clotrimazole condition. This cluster represents the fact that no significant changed occurred in the first 30 minutes. The Clotrimazole is clustered with this 'null change' cluster because there was no significant protein abundance change in any of the genes in this condition. The stress genes, from time point 3 and above are clustered together (green cluster) and sub-cluster to their conditions. Overall this clustering is according to expectation, divides the condition accurately and teaches us that 30 minutes are not enough for protein abundance significant change in response to environmental perturbations. Only from

the third time point, which corresponds to one hour, there seems to be a noticeable effect of the condition.

Using the correlation between the noise residuals we get the dendogram in figure 18.

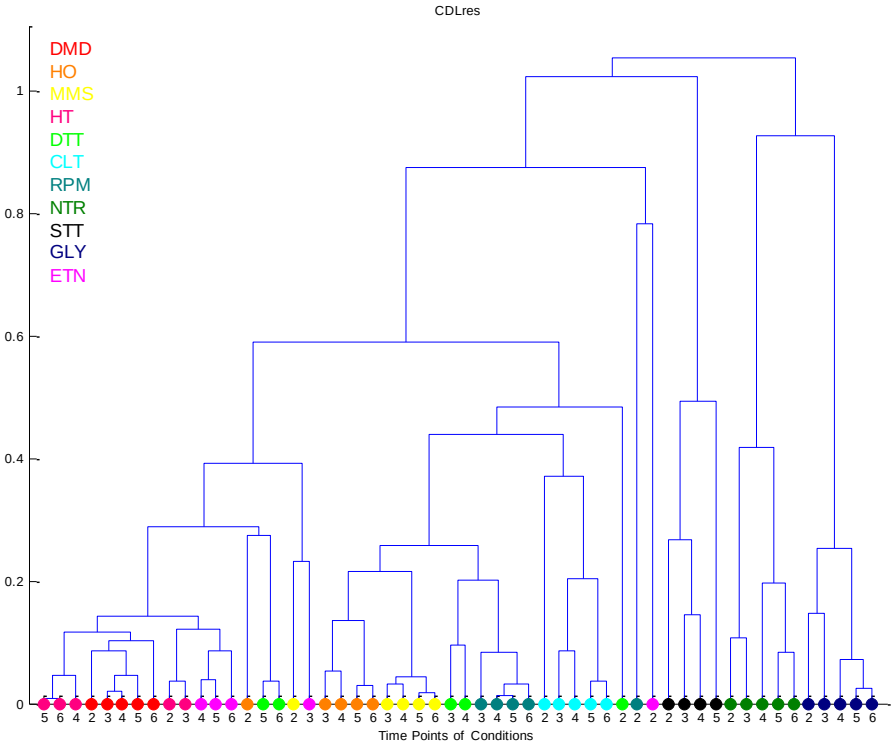

Figure 18: Dendogram of the time points and conditions using Pearson Correlation on the noise values and clustering by average linkage method. Numbers and coloring as in figure 17.

As in the clustering of the relative mean values the time points of specific conditions are generally clustered together. The differentiation between the stress and the stress relaxation conditions remains significant. The early time point's cluster seems to break to two sub-clusters, as the late time points, although the early ones are still distinct. Here again, clotrimazole seems not to produce much of noise residuals change.

Clustering of genes in conditions

In order to get insight on the relative effect of module vs. condition affiliation we preformed a clustering of the genes and the conditions together. We created a list of 38·11 vectors $v_{G,C}$. Each such vector contains the five relative mean values corresponding to all the time points (but the first, which we used to normalize the relative values) of gene G in condition C. We clustered those vectors using the average linkage clustering methods. The dendogram we got has a strong division into two clusters (not shown). The separation of the vectors into those clusters is shown in figure 19.

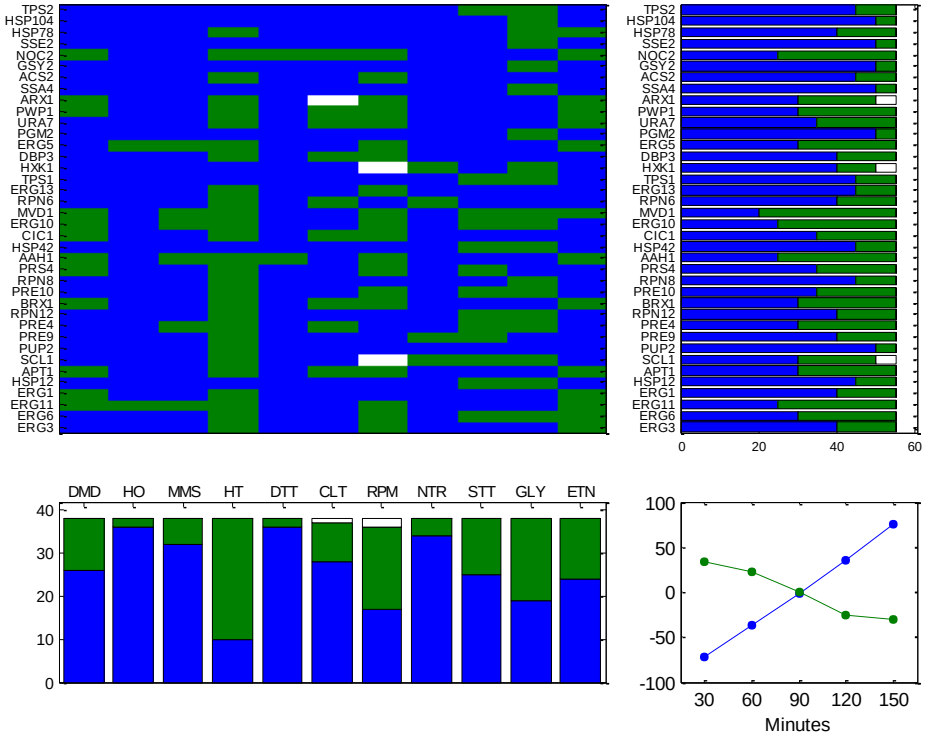

Figure 19: Upper left. The division of the 'gene-in-condition' time points vectors to the two distinct clusters, using the relative mean values. Green bars – vectors that belong to the first cluster, blue bars – vectors that belong to the second cluster (white bars correspond to no data – experimental failures). **Upper right**. Histogram according to the genes. For each gene, the number of conditions in which its pattern clustered with each profile. **Bottom left**. Histogram according to the conditions. For each condition, the number of genes that their pattern clustered with each profile. **Bottom right**. The normalized and centered average profile of each cluster.

As seen from the bottom right sub figure, the two clusters have unique profiles of change in time, one increasing, while the other decreasing. Both the module and the condition affiliation have an important role in determining to which cluster each vector will belong to. For example, stress genes tend to be enriched in the increasing

cluster while the decreasing cluster contains primarily ergosterol and rRNA processing genes. It is interesting to note the heat shock tends to produce a decreasing profile, in contrary to most other stress conditions (apart from the stress genes that are still induced under that condition).

The same analysis was preformed using the noise values. Here we got five distinct clusters, as shown is figure 20.

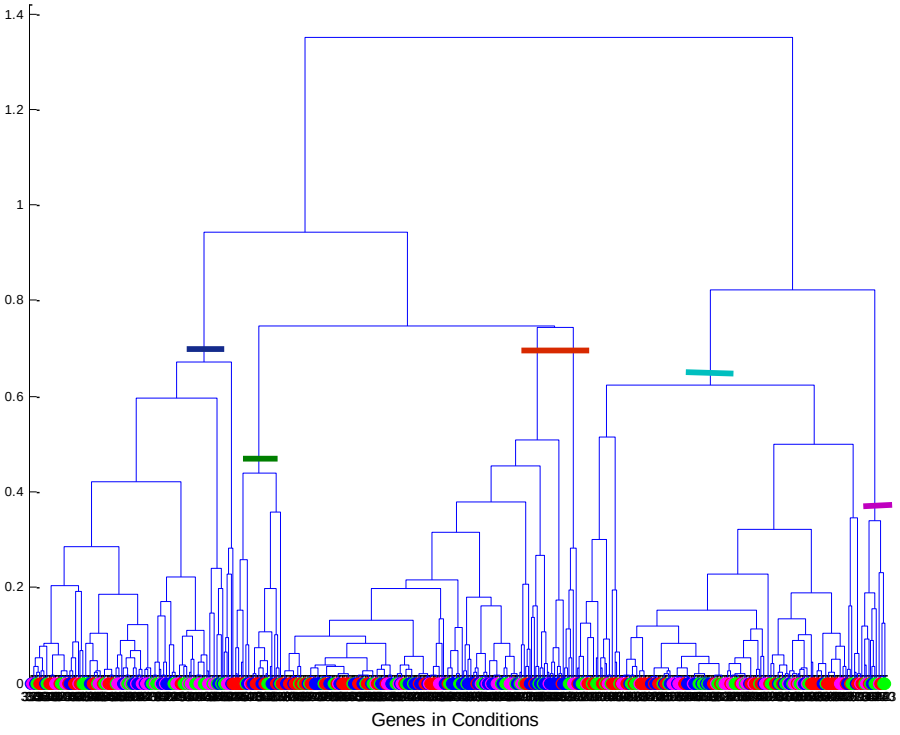

Figure 20. Dendogram of the genes in conditions using Pearson Correlation on the relative noise residuals values and clustering by average linkage method. The common roots of the 5 clusters are marked with colored horizontal line.

The separation of the vectors into those clusters is shown in figure 21. Of the five clusters, three present interesting, non-overlapping profiles in time. Those profiles are displayed by all genes, at some conditions, and in almost all conditions, by some genes. Module affiliation seems to play a lesser role than condition. It is interesting to note that in two conditions, glycerol growth relaxation and ethanol, there is only one dominant noise residuals profile, which is probably indicative to the effect of that conditions.

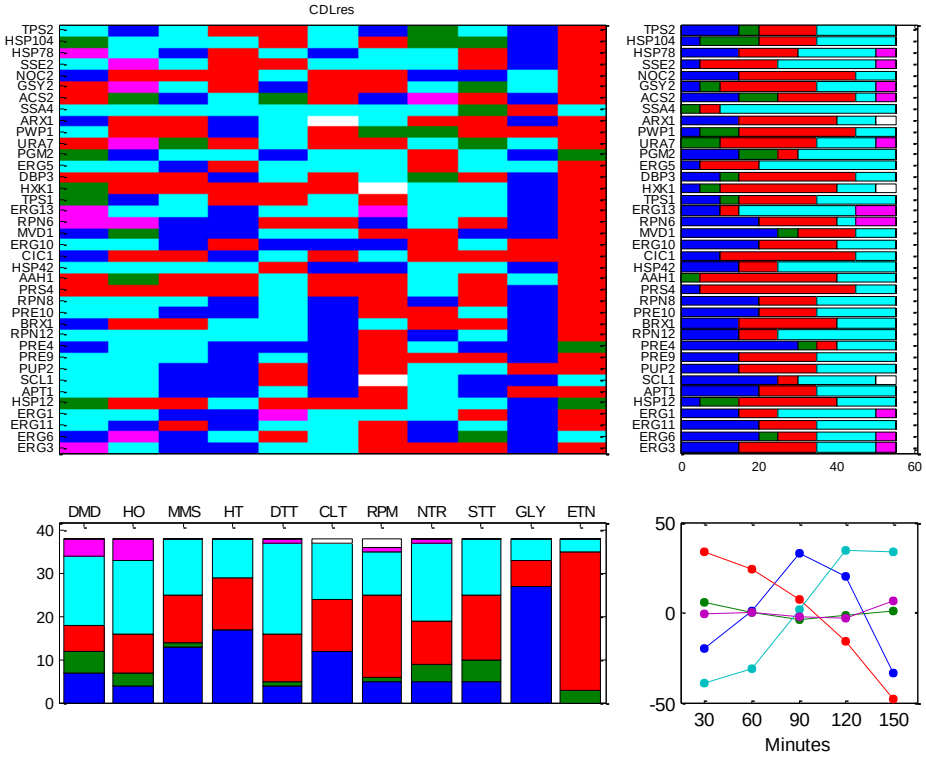

Figure 21: Upper left. The division of the 'gene-in-condition' time points vectors to the five distinct clusters, using the noise values. Colors represent the cluster root in the dendogram presented in figure 20. White bars correspond to no data. **Upper right**, **Bottom left** and **Bottom right** as in figure 19.

Noise residuals and promoter motifs

The importance of promoter motifs and transcription factors (TFs) for controlling the transcription rate and hence the mean protein abundance is well established. We wanted the check whether they have an influence on the noise residuals as well. Each of our chosen modules has few TFs binding to several of the chosen genes belonging to it. Table 5 summarizes those TFs/regulatory motifs (PAC and mRRPE are not yet known to be bound by TFs).

In order to quantify the effect of those motifs on the mean and noise of the protein abundance we created a novel definition of expression coherence (EC). The EC score is a measurement of the extent to which the correlation between the set of genes having the motif is higher than the correlation between those genes and other, which do not have that motif (see Methods). The EC score was calculated for the above motifs using the mean, relative mean, noise residuals and relative noise residuals values. The results are summarized in figure 22.

TF	Regulated Module (total genes from our chosen genes)	Form our 38 analyzed genes, number of regulated genes form:	
		The regulated module	Other modules
HSE	Stress (11)	5	0
MSN4	Stress (11)	5	0
HSF1	Stress (11)	4	0
MSN2	Stress (11)	5	1
RPN4	Proteasome (8)	7	0
HAP1	Ergosterol (9)	5	0
PAC	rRNA Processing (10)	7	0
mRRPE	rRNA Processing (10)	7	2

Table 5: Summary of the major TFs/regulatory motifs regulating the different modules.

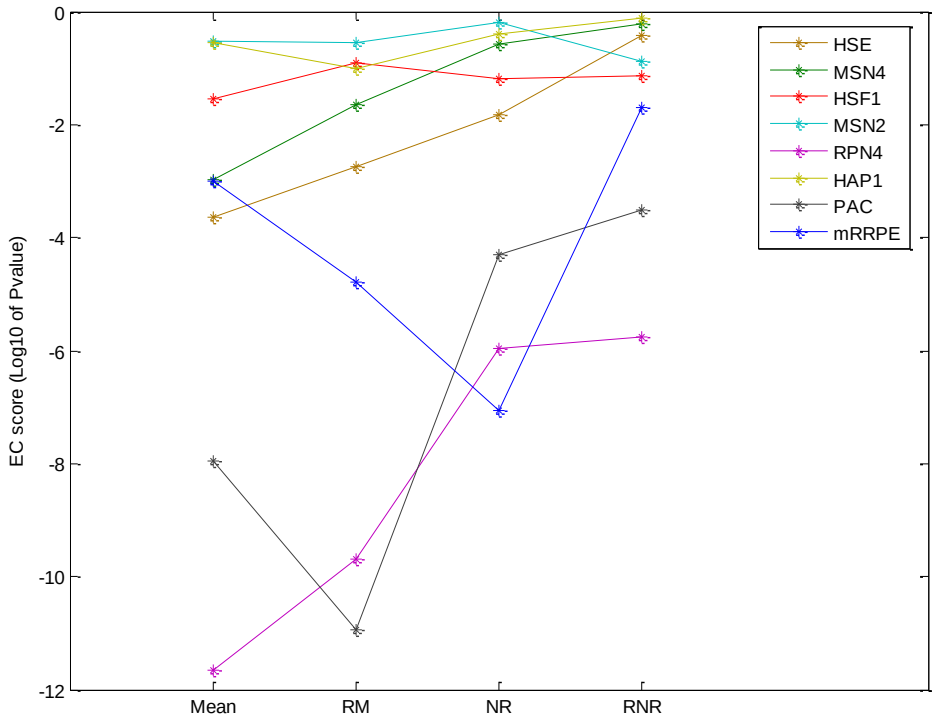

Figure 22: EC score (see text and Methods) of the major motifs using the mean, the relative mean (RM), the noise residuals (NR) and the relative noise residuals (RNR) values. Lower EC score mean higher relative correlation of the genes containing the motif.

Most motifs have better EC score for the mean values than for the noise residuals, which suggest that transcription regulation has less significance in determining the noise than in influencing the mean protein abundance. One interesting expectation is

mRRPE, which has lower (i.e. more significant) EC score using the noise residuals, although this trend does not hold using the relative noise residuals.

DISCUSSION

We have tested the mean and the noise of protein abundance of 43 different *S. cerevisiae* genes, using a fused GFP reporter. The genes belong to 4 distinct expression modules – stress, proteasome, ergosterol and rRNA processing. The fluorescence distributions were obtained for 11 different environmental perturbations, each measured in six consecutive time points.

The shapes of the fluorescence distributions present interesting patterns. We found a strong correlation between the mean abundance of a protein and its distribution's resemblance to normal or log-normal distribution. High abundant proteins are characterized in log-normal distributions, while low abundant ones display normal distributions. The origin of both types of distributions is different: normal distributions are usually created by the summation of random variables (the central limit theorem), while log-normal distribution can originate from the multiplication of such variables. However, there is no apparent reason for why high abundant protein will be a product of variables, while low abundant ones will be their summation. The background fluorescence displays a clear normal distribution, and its effect might play a non-negligible role at determining the fluorescence pattern of the proteins with the lowest abundance, although it is not reasonable to assume it has an effect on protein with mid-range and high abundance. Most stress genes are characterized with relative high skewness of linear fluorescence values, as compared to the skewness of genes from other modules and as expected by their mean expression (even mediocre express stress genes have a relative high skewness). This finding could indicate the existence of some unique, non-Gaussian, variable(s) or factor(s) influencing the expression of those proteins and shifting their distributions more to the log-normal pattern.

The observed mean abundances of the proteins agreed both with the published mRNA data and with the module affiliation of the tested genes. The deviation from the accordance could be explained by experimental noise, originate from the relatively high background fluorescence, or biological mechanism such as post-transcriptional regulation. Much of the clustering inconsistency with the module affiliation can be attributed to the agglomerative hierarchical nature of the average linkage clustering method. Indeed, the PCA show a very nice elongated structure, on which the modules

are ordered consecutively in accordance with their expected reaction to stressful conditions (stress genes being highly induces, proteasome mildly induced, ergosterol mildly repressed and rRNA processing highly repressed).

The observed noise in protein abundance presents several strong dependencies on the mean abundance, at different regimes. While the $\sigma^2/\mu^2 \sim 1/\mu^2$ dependency was easy to decipher, as a trend that originates from the non-negligible background fluorescence, the others were more elusive. The $\sigma^2/\mu^2 \sim 1/\mu$ dependency was consistent in all conditions and time points and covering the regime of the low proteins abundance (that were still above the background fluorescence). Our interpretation for this dependency as the contribution of low abundant mRNA seems to fit the data in the best way. However, experimental framework of previous publications found that the measured intrinsic noise (which includes the mRNA noise, in their experimental set-up) has only a negligible contribution to the overall noise. This is a central discrepancy. This inconsistency is even more problematic if considering two works that showed that the dominant source of stochasticity comes from global factors. Raser et al.(23) demonstrated a very high correlation between the abundance of reporters, derived from different promoters, and Colman-Lerner et al.(20) separated pathway related from global factors and discovered that the latter ones have the primary role. Yet, our analytical analysis clearly shows that global contributors will not create a $\sigma^2/\mu^2 \sim 1/\mu$ dependency, but rather a $\sigma^2/\mu^2 \sim C$ dependency. A possible resolution of the discrepancy is that many of the published experiments were done in high abundant proteins, which are indeed located at the $\sigma^2/\mu^2 \sim C$ regime that fit the global dominance hypothesis. However, some of the measured reporters were relatively low abundant, and derived from week promoters and still have low 'intrinsic' noise. The final conclusion of this issue will have to be postponed to latter experiments.

The $\sigma^2/\mu^2 \sim C$ dependency that is observed for high abundant proteins may come from low abundant global factor. However, there is no such trivial factor, which is present in the right number of copies. Alternatively un-equal cell division can create the $\sigma^2/\mu^2 \sim C$ dependency. The calculated partition coefficient mean and STD values that mach the measured global extrinsic contribution are indeed in the reasonable

biological range – a mean partition coefficient between 0.4 to 0.5 and a STD between 0 to 0.1.

Many genes have noise level much higher than expected by the trend lines, giving their mean. We termed those deviations noise residuals. The noise residuals are probably composed from several components; some are module related, while others are gene specific, such as chromatin arrangement near the gene. We have showed that module affiliation, although not responsible for the primary effect on the general trends, plays a dominant role in determining the noise residuals. Stress genes are highly significant noisier than others, while proteasome genes are characterized in small noise residuals.

The most reasonable mechanism by which module affiliation affects stochasticity in protein abundance is the common regulators, shared by the module genes. MSN2 and MSN4, the common regulators of the stress module are indeed thought to be low abundant. MSN4, especially, is considered to be present in few dozens copies only. Because both those transcription factors regulate hundreds of genes, there are too little copies of them to occupy all promoters. Hence, MSN4 and MSN2 might 'jump' from one promoter to other, creating an additional noise in gene expression.

Other factors that can be important to the module related noise could be a shared selective pressure applied to all the module genes. For example, Fraser et al.(25) hypothesized that essential genes and genes that work in a complex will be characterized in low noise level. Our data do support those evolutionary considerations – proteasome genes, which work in a complex and are all, but on, essential are the with smallest noise residuals, while the stress genes, which are all dispensable, have the highest noise residuals. Two of the stress genes: TPS1 and TPS2 (circled in figure 13) also work in a complex. Their noise residuals (8 and 9.4, respectively) are among the less noisy half of the stress genes, thought they are not with the lowest noise residuals among them. An intriguing possibility is that for some genes (as the stress genes), in some conditions, enhanced noise may even be beneficial at the population level. It remains to be understood whether the enhanced noise observed here in stress genes could have been selected for by evolution or is it a mere result of lack of constraint on the expression of such genes.

Because module affiliation indicates, in many situations, the upstream motifs composition of a gene, we tried to correlate their presence with the noise residuals. Generally, the effect of motifs on noise residuals coherence was much less profound than there effect of protein abundance coherence. One rRNA processing motif, mRRPE, seems to work better on noise residuals than on mean abundance, but this finding can be a result of the noisy data originating from the small number of genes the motif is regulating. It will be interesting to try and seek out noise specialized motifs, but it will require much larger gene set.

Chromatin remodeling was expected to play a significant role in determining the noise level, as indicated by several works(21-23). However, we did not find any indications that strengthen that hypothesis when we used both experimental and computational data regarding chromatin properties. Even if this effect do exists, it was overshadowed by the module affiliation. A nice example to this pattern is given by the genes HXK1, RPN12 and PRE4. The distance between the stress gene HXK1 to the proteasome genes RPN12 and PRE4 is only 263 and 5183 base pairs, respectively; so it is reasonable they share roughly the same chromatin properties. Yet, HXK1 is one of the highly fluctuating proteins, while RPN12 and PRE4 have very small noise residuals. It is possible that having more genes to analyze would reveal a correlation that is hidden in the relatively small gene set, comprising our current research.

Clustering the time points of the different conditions according the mean abundance of the noise residuals reveals several patterns. First, a clear separation is observed between the stress conditions and the stress relaxation ones. Second, the clustering together of first time points of all conditions, for both mean and noise residuals values, is an indication that a significant genetic response to perturbations does not occur in the first 30 minutes of the experiment. Lastly, one condition, Clotrimazole, cluster together with the first time points, which is a sign that it has only minor influence on protein abundance – both mean and noise.

Analytical calculations, together with some experimental works have claimed that noise level should arise when the protein level is not in its steady state(s). The protein abundance in our experiments can be viewed as a system that is going through a

disruption – leaving an old steady state and stabilizing on a new one. Therefore we had expected the average profile of noise level across time to have an inverted U shape – initially noise will increase, which corresponds to exiting the first steady state, and than it will decrease, after reaching the second steady state. Fitting the noise data to the mean abundance, from all the genes, for each time point of each condition showed no significant change of the trend line between early and late time points. Moreover, the noise residuals do not have single response pattern to perturbations. On the contrary, the response pattern changed dramatically both between the different modules and across the conditions. The entire possible patterns spectrum was observed – monotonically decreasing residuals, monotonic increasing residual, U-shape behavior, inverted-U-shape behavior and inconsistent pattern. This finding is even more interesting because, in most cases, there was a pattern consistency among genes of the same module, in the same condition. High consistency patterns within condition and module, but very low one across conditions or modules is indeed a non trivial finding. More research effort should be invested in to this direction in the future.

REFERENCES

1. McAdams, H. H. & Arkin, A. (1999) *Trends Genet* **15,** 65-9.
2. McAdams, H. H. & Arkin, A. (1997) *Proc Natl Acad Sci U S A* **94,** 814-9.
3. Elowitz, M. B. & Leibler, S. (2000) *Nature* **403,** 335-8.
4. Barkai, N. & Leibler, S. (2000) *Nature* **403,** 267-8.
5. Berg, O. G., Paulsson, J. & Ehrenberg, M. (2000) *Biophys J* **79,** 1228-36.
6. Rao, C. V., Wolf, D. M. & Arkin, A. P. (2002) *Nature* **420,** 231-7.
7. Paulsson, J. (2004) *Nature* **427,** 415-8.
8. Spudich, J. L. & Koshland, D. E., Jr. (1976) *Nature* **262,** 467-71.
9. Maloney, P. C. & Rotman, B. (1973) *J Mol Biol* **73,** 77-91.
10. Lobner-Olesen, A. (1999) *Embo J* **18,** 1712-21.
11. Becskei, A. & Serrano, L. (2000) *Nature* **405,** 590-3.
12. Arkin, A., Ross, J. & McAdams, H. H. (1998) *Genetics* **149,** 1633-48.
13. Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, Y. & Sakano, H. (2003) *Science* **302,** 2088-94.
14. van de Putte, P. & Goosen, N. (1992) *Trends Genet* **8,** 457-62.
15. von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. (2000) *Nature* **406,** 188-92.
16. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. (2002) *Nat Genet* **31,** 69-73.
17. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002) *Science* **297,** 1183-6.
18. Thattai, M. & van Oudenaarden, A. (2001) *Proc Natl Acad Sci U S A* **98,** 8614-9.
19. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. (2005) *Science* **307,** 1962-5.
20. Colman-Lerner, A., Gordon, A., Serra, E., Chin, T., Resnekov, O., Endy, D., Pesce, C. G. & Brent, R. (2005) *Nature* **437,** 699-706.
21. Becskei, A., Kaufmann, B. B. & van Oudenaarden, A. (2005) *Nat Genet* **37,** 937-44.
22. Blake, W. J., M, K. A., Cantor, C. R. & Collins, J. J. (2003) *Nature* **422,** 633-7.
23. Raser, J. M. & O'Shea, E. K. (2004) *Science* **304,** 1811-4.
24. Pedraza, J. M. & van Oudenaarden, A. (2005) *Science* **307,** 1965-9.
25. Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. (2004) *PLoS Biol* **2,** e137.
26. Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. (2000) *Proc Natl Acad Sci U S A* **97,** 2075-80.
27. Kepler, T. B. & Elston, T. C. (2001) *Biophys J* **81,** 3116-36.
28. Eldar, A. & Elowitz, M. (2005) *Nature* **437,** 631-2.
29. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. & Weissman, J. S. (2003) *Nature* **425,** 737-41.
30. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) *Mol Biol Cell* **11,** 4241-57.
31. Greenbaum, D., Jansen, R. & Gerstein, M. (2002) *Bioinformatics* **18,** 585-96.
32. Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. & Brown, P. O. (2002) *Proc Natl Acad Sci U S A* **99,** 5860-5.

33. Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K. & Young, R. A. (2005) *Cell* **122,** 517-27.
34. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399-403.
35. dos Reis, M., Savva, R. & Wernisch, L. (2004) *Nucleic Acids Res* **32,** 5036-44.
36. Kafri, R., Bar-Even, A. & Pilpel, Y. (2005) *Nat Genet* **37,** 295-9.
37. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. (2004) *Cell* **116,** 699-709.
38. Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T., Chu, A. M., Giaever, G., Prokisch, H., Oefner, P. J. & Davis, R. W. (2002) *Nat Genet* **31,** 400-4.