TITLE: Selection at the pathway level drives the evolution of gene-specific transcriptional noise

AUTHORS: Gustavo Valadares Barroso[1]; Natasa Puzovic[1] and Julien Y Dutheil[1,2]

Affiliations:

1) Max Planck Institute for Evolutionary Biology. Department of Evolutionary Genetics. August-Thienemann-Straße 2 24306 Plön – GERMANY

2) ISEM – Institut des Sciences de l'Évolution. UMR 5554, Université de Montpellier, Place Eugène Bataillon 34095 Montpellier cedex 05 – FRANCE

Corresponding Author:

Gustavo V. Barroso, Max Planck Institute for Evolutionary Biology. Department of Evolutionary Genetics. August-Thienemann-Straße 2, 24306 Plön – GERMANY.

ABSTRACT:

Biochemical reactions within individual cells result from the interactions of molecules, often in small numbers. Consequently, the inherent stochasticity of binding and diffusion processes generate noise along the cascade that leads to the synthesis of a protein from its encoding gene. As a result, isogenic cell populations display phenotypic variability even in homogeneous environments. The extent and consequences of this stochastic gene expression have only recently been assessed on a genome-wide scale, in particular owing to the advent of single cell transcriptomics. However, the evolutionary forces shaping this stochasticity have yet to be unraveled. We took advantage of two recently published data sets of the single-cell transcriptome of the domestic mouse *Mus musculus* in order to characterize the effect of natural selection on gene-specific transcriptional stochasticity. We showed that noise levels in the mRNA distributions (*a.k.a.* transcriptional noise) significantly correlate with three-dimensional nuclear domain organization, evolutionary constraint on the encoded protein and gene age. The position of the encoded protein in biological pathways, however, is the main factor that explains observed levels of transcriptional noise, in agreement with models of noise propagation within gene networks. Because transcriptional noise is under widespread selection, we argue that it constitutes an important component of the phenotype and that variance of expression is a potential target of adaptation. Stochastic gene expression should therefore be considered together with mean expression level in functional and evolutionary studies of gene expression.

Introduction

Isogenic cell populations display phenotypic variability even in homogeneous environments (Spudich and Koshland 1976). This observation challenged the clockwork view of the intra-cellular molecular machinery and led to the recognition of the stochastic nature of gene expression. Because biochemical reactions result from the interactions of individual molecules in small numbers (Gillesple 1977), the inherent stochasticity of binding and diffusion processes generates noise along the biochemical cascade leading to the synthesis of a protein from its encoding gene (**Figure 1**). The study of stochastic gene expression (SGE) classically recognizes two sources of expression noise. Following the definition introduced by Elowitz et al (Elowitz et al. 2002), extrinsic noise results from variation in concentration, state and location of shared key molecules involved in the reaction cascade from transcription initiation to protein folding. This is because molecules that are shared among genes, such as ribosomes and RNA polymerases, are typically present in low copy numbers relative to the number of genes actively transcribed (Shahrezaei and Swain 2008). Extrinsic factors also include physical properties of the cell such as size and growth rate, likely to impact the diffusion process of all molecular players. Extrinsic factors therefore affect every gene in a cell equally. Conversely, intrinsic factors generate noise in a gene-specific manner. They involve, for example, the strength of cis-regulatory elements (Suter et al. 2011) as well as the stability of the mRNA molecules that are transcribed (Mcadams and Arkin 1997; Thattai and Oudenaarden 2001). Every gene is affected by both sources of stochasticity and the relative importance of each has been discussed in the literature (Becskei et al. 2005; Raj and Oudenaarden 2008). Shahrezaei and Swain (Shahrezaei and Swain 2008) proposed a more general, systemic and explicit definition for any organization level, where intrinsic stochasticity is “generated by the dynamics of the system from the random timing of individual reactions” and extrinsic stochasticity is “generated by the system interacting with other stochastic systems in the cell or its environment”. This generic definition therefore includes Raser and O’Shea's (Raser and O’Shea 2005) suggestion to further distinguish extrinsic noise occurring “within pathways” and “between pathways”. Other organization levels of gene expression are also likely to affect expression noise, such as chromatin structure (Blake et al. 2003; Hebenstreit 2013), and three-dimensional genome organization (Pombo and Dillon 2015).

Pioneering work by Fraser et al (Fraser et al. 2004) has shown that SGE is an evolvable trait which is subject to natural selection. First, genes involved in core functions of the cell are expected to behave more deterministically (Barkai and Leibler 1999) because temporal oscillations in the concentration of their encoded proteins are likely to have a deleterious effect. Second, genes involved in immune response (Arkin et al. 1998; Norman et al. 2015) and response to environmental conditions can benefit from being unpredictably expressed in the context of selection for bet-hedging (Thattai and Oudenaarden 2004). As the relation between fitness and stochasticity depends on the function of the underlying gene, selection on SGE is expected to act mostly at the intrinsic level (Newman et al. 2006; Lehner 2008; Wang and Zhang 2011). The molecular mechanisms by which natural selection operates to regulate expression noise, however, remain to be elucidated.

Due to methodological limitations, seminal studies on SGE (both at the mRNA and protein levels) have focused on only a handful of genes (Elowitz et al. 2002; Ozbudak et al. 2002; Chubb et al. 2006). The canonical approach consists in selecting genes of interest and recording the change of their noise levels in a population of clonal cells as a function of either (1) the concentration of the molecule that allosterically controls affinity of the transcription factor to the promoter region of the gene (Blake et al. 2003; Bar-even et al. 2006) or (2) mutations artificially imposed in regulatory sequences (Ozbudak et al. 2002). In parallel with theoretical work (Kepler and Elston 2001; Kaufmann and van Oudenaarden 2007; Sánchez and Kondev 2008), these pioneering studies have provided the basis of our current understanding of the proximate molecular mechanisms behind SGE, namely complex regulation by transcription factors, architecture of the upstream region (including the presence of TATA box), translation efficiency and mRNA / protein stability (Eldar and Elowitz 2010). Measurements at the genome scale are however needed in order to go beyond gene idiosyncrasies and particular histories, and test hypotheses about the evolutionary forces shaping SGE (Sauer et al. 2007).

The recent advent of single-cell RNA sequencing makes it possible to sequence the transcriptome of each individual cell in a collection of clones, and to observe the variation of gene-specific mRNA quantities across cells. This provides a genome-wide assessment of transcriptional noise. While not accounting for putative noise resulting from the process of translation of mRNAs into proteins, transcriptional noise accounts for noise generated by both synthesis and degradation of mRNA molecules (**Figure 1**). Previous studies, however, have shown that transcription is a limiting step in gene expression, and that transcriptional noise is therefore a good proxy for expression noise (Newman et al. 2006; Taniguchi et al. 2011). Here, we used publicly available single-cell transcriptomics data sets to quantify gene-specific transcriptional noise and relate it to other genomic factors, including protein conservation and position in the interaction network, in order to uncover the molecular basis of selection on stochastic gene expression.

Results

A new measure of noise to study genome-wide patterns of stochastic gene expression

We used the dataset generated by Sasagawa et al (2013), which quantifies gene-specific amounts of mRNA as fragments per kilobase of transcripts per million mapped fragments (FPKM) values for each gene and each individual cell. Among these, we selected all genes in a subset containing 20 embryonic stem cells in G1 phase in order to avoid recording variance that is due to different cell types or cell-cycle phases. The Quartz-Seq sequencing protocol captures every poly-A RNA present in the cell at one specific moment, allowing to assess transcriptional noise. Following Shalek et al (2014) we first filtered out genes that were not appreciably expressed in order to reduce the contribution of technical noise to the total noise. For each gene we further calculated the mean μ in FPKM units and variance σ^2 in FPKM^2 units, as well as two previously published measures of stochasticity: the *Fano factor*, usually referred to as the bursty parameter, defined as σ^2/μ and *Noise*, defined as the coefficient of variation squared (σ^2/μ^2). Both the variance and *Fano factor* are monotonically increasing functions of the mean (**Figure 2A**). *Noise* is inversely proportional to mean expression (**Figure 2A**), in agreement with previous observations at the protein level (Bar-even et al. 2006; Taniguchi et al. 2011). While this negative correlation was theoretically predicted (Tao et al. 2007), it may confound the analyses of transcriptional noise at the genome level, because mean gene expression is under specific selective pressure (Pál et al. 2001). In order to disentangle these effects, we developed a new quantitative measure of noise, independent of the mean expression level of each gene. To achieve this we fitted a linear model in the log-space plot of variance *versus* mean and extracted the slope (a) and intercept (b) of the regression line. We defined F* as $\sigma^2/(a.\mu^b)$ (see Material and Methods) that is, the ratio of the observed variance over the variance component predicted by the mean expression level. Genes with F* < 1 have a variance lower than expected according to their mean expression whereas genes with F* > 1 behave the opposite way **(Figure 2A**). This approach is similar in principle to the running median approach of Newmann et. al (Newman et al. 2006). As expected, F* displays no significant correlation with the mean (Kendall's tau = -0.009, p-value = 0.106**, Figure 2B**). We therefore use F* as a measure of SGE throughout this study.

Stochastic gene expression correlates with the three-dimensional, but not one-dimensional, structure of the genome

We first sought to investigate whether genome organization significantly impacts the patterns of stochastic gene expression. We assessed whether genes in proximity along chromosomes display more similar amount of transcriptional noise than distant genes. We tested this hypothesis by computing the primary distance on the genome between each pair of genes, that is, the number of base pairs separating them on the chromosome, as well as the relative difference in their transcriptional noise (see Methods). We found no significant association between the two distances (Mantel tests, each chromosome tested independently). Contiguous genes in one dimension, however, have significantly more similar transcriptional noise that non-contiguous genes (permutation test, p-value < 1e-3, **Figure S1**). Using Hi-C data from mouse embryonic cells (Dixon et al. 2012), we report that genes in contact in three-dimensions have significantly more similar transcriptional noise than genes not in contact (permutation test, p-value < 1e-3, **Figure S1**). Most contiguous genes in one-dimension also appear to be close in three-dimensions and the effect of 3D contact is stronger than that of 1D contact**.** These results therefore suggest that the three-dimensional structure of the genome has a stronger impact on stochastic gene expression than the position of the genes along the chromosomes. We further note that while highly significant, the size of this effect is small, with a difference in relative expression of -1.12% (**Figure S1**).

Transcription factors binding and histone methylation impact stochastic gene expression

The binding of transcription factors (TF) to promoter constitutes one notable source of transcriptional noise (**Figure 1**) (Blake et al. 2003; Newman et al. 2006). In eukaryotes, the accessibility of promoters is determined by the chromatin state, which is itself controlled by histone methylation. We assessed the extent to which transcriptional noise is linked to particular TFs and histone marks by using data from the Ensembl regulatory build (Zerbino et al. 2015), which provides data from experimental evidence of TF binding and methylation sites along the genome. First we contrasted the F* values of genes with binding evidence for each annotated TF independently. Among 13 TF represented by at least 5 genes in our data set, we found that 4 of them significantly influence F* after adjusting for a global false discovery rate of 5%: the transcription repressor CTFC (adjusted p-value = 0.0286), the transcription factor CP2-like 1 (Tcfcp2l1, adjusted p-value = 0.0111), the X-Linked Zinc Finger Protein (Zfx, adjusted p-value = 0.0111) and the Myc transcription factor (MYC, ajusted p-value = 0.0111). Interestingly, association with each of these four TFs led to an increase in transcriptional noise. We also report a weak but significant positive correlation between the number of transcription factors associated with each gene and the amount of transcriptional noise (Kendall's tau = 0.023, p-value = 0.0009). This observation is consistent with the idea that noise generated by each TF is cumulative(Sharon et al. 2014). We then tested if particular histone marks are associated with transcriptional noise. Among five histone marks represented in our data set, three were found to be highly significantly associated to a higher transcriptional noise: H3K4me3 (adjusted p-value = 3.032e-162), H3K4me2 (adjusted p-value = 1.01e-129) and H3K27me3 (adjusted p-value = 7.418e-33). Methylation on the fourth Lysine of histone H3 is associated with gene activation in humans, while tri-methylation on lysine 27 is usually associated with gene repression (Barski et al. 2007). These results suggest that both gene activation and silencing contribute to the stochasticity of gene expression, in agreement with the view that bursty transcription leads to increased noise (Blake et al. 2003; Newman et al. 2006).

Low noise genes are enriched for housekeeping functions

We investigated the function of genes at both ends of the F* spectrum. We defined as candidate gene sets the top 10% least noisy or the top 10% most noisy genes in our data set, and tested for enrichment of GO terms and Reactome pathways (see Methods). It is expected that genes encoding proteins participating in housekeeping pathways are less noisy because fluctuations in concentration of their products might have stronger deleterious effects (Pedraza and van Oudenaarden 2005). On the other hand, stochastic gene expression could be selectively advantageous for genes involved in immune and stress response, as part of a bet-hedging strategy (eg Arkin et al. 1998; Shalek et al. 2013). While we do not find any significantly enriched Reactome pathway in the high noise gene set, a total of 37 pathways were significantly over-represented in the low-noise gene set (false discovery rate set to 1%). Interestingly, the top most significant pathways belong to modules related to translation (initiation, elongation, termination as well as ribosomal assembly), as well as several modules relating to gene expression, including chromatin regulation and mRNA splicing (**Figure 3**). GO terms enrichment tests lead to similar results (**Table 1**): we found the molecular functions "nucleic acid binding" and "structural constituent of ribosome", the biological processes "nucleosome assembly", "innate immune response in mucosa" and "translation", as well as the cellular component "nuclear nucleosome" to be enriched in the low noise gene set. All these terms but one relate to gene expression.

The lack of significantly enriched Reactome pathways by high noise genes can potentially be explained by the nature of the data set: as the original experiment was based on unstimulated cells, genes that directly benefit from high SGE might not be expressed in these experimental conditions. In accordance, high-noise genes are not found to be enriched for any GO term.

Highly connected proteins are synthesized by low-noise genes

The structure of the interaction network of proteins inside the cell can greatly impact the evolutionary dynamics of genes (Jeong et al. 2000; Barabási and Oltvai 2004). Furthermore, the contribution of each constitutive node within a given network varies. This asymmetry is largely reflected in the power-law-like degree distribution that is observed in virtually all biological networks (Barabási and Albert 1999) with a few genes displaying a lot of connections and a majority of genes displaying only a few. The individual characteristics of each node in a network can be characterized by various measures of centrality (Newmann 2003). Following previous studies on protein evolutionary rate (Fraser et al. 2002; Hahn et al. 2004; Jovelin and Phillips 2009) we asked whether, at the gene level, there is a link between centrality of a protein and the amount of transcriptional noise as measured by F*, using five centrality metrics measured from the pathway data available in the Reactome database (Croft et al. 2014). Our data set encompasses 13,660 genes for which both gene expression data and pathway annotations were available.

We first estimated the pleiotropy index of each gene by counting how many different pathways the corresponding proteins are involved in. We then computed centrality measures as averages over all pathways in which each gene is involved. These measures include (1) *node degree* (here simply referred to as “degree”), which corresponds to the number of other nodes a given node is directly connected with, (2) *hub score*, which estimates the extent to which a node links to other central nodes, (3) *authority score*, which estimates the importance of a node by assessing how many hubs link to it, (4) *closeness*, a measure of the topological distance between a node and every other reachable node (the fewer edge hops it takes for a protein to reach every other protein in a network, the higher its closeness), and (5) *betweenness*, a measure of the frequency with which a protein belongs to the shortest path between every pair of nodes.

A principal component analysis (PCA) revealed that these measures essentially fall in two groups (**Figure S2**). The first component explained 43.4% of the total inertia, and represents measures relating to the number of interacting partners of a given protein (degree 31.9%, hub score 32.8%, authority score 33.6%). The second component, explaining 17.5% of the total inertia, represents the other three variables (pleiotropy 41.3%, betweeness 15.7%, closeness 40.6%). The third axis (17.2% of total inertia) represents only two variables (betweenness, 59.3% and closeness 38.4%), while the fourth axis (15.3% of total inertia) represents in majority pleiotropy (54.8%).

Measures contributing to the first component of the PCA are all significantly negatively correlated with transcriptional noise: the more central a protein is, the less transcriptional noise it displays (**Table 2**). We also observed that pleiotropy is negatively correlated with F* (**Table 2**), although to a lesser extent suggesting that a protein that potentially performs multiple functions at the same time needs to be less noisy. This effect is not an artifact of the fact that pleiotropic genes are themselves more central (e.g. correlation of pleiotropy and node degree: Kendall’s tau = 0.229, p-value < 2.2e-16) or evolve more slowly (correlation of pleiotropy and Ka / Ks ratio: Kendall’s tau = -0.11, p-value < 2.2e-16) since it is still significant after controlling for these variables (partial correlation of pleiotropy and F*, accounting for centrality measures and Ka / Ks: Kendall's tau = -0.036, p-value = 3.695e-10). Closeness and betweenness, on the other hand, are highly correlated with each other but are independent of the degree measures (**Figure S2**), and do not significantly correlate with F* (**Table 2**). In modular networks (Hartwell et al. 1999) nodes that connect different modules are extremely important to the cell (Guimera and Amaral 2005) and show high betweenness scores. In yeast, high betweenness proteins tend to be older and more essential (Joy et al. 2005), an observation also supported by our data set (betweenness *vs* gene age, Kendall's tau = 0.077, p-value = 7.569e-10; betweenness *vs* Ka/Ks, Kendall's tau = -0.077, p-value = 7.818e-12). It has been argued, however, that in protein-protein interaction networks high betweenness proteins are less essential due to the lack of directed information flow, compared to, for instance, regulatory networks (Yu et al. 2007), a hypothesis which could explain the lack of observed correlation.

It was previously shown that centrality measures negatively correlates with evolutionary rate (Hahn and Kern 2004). Our results suggest that central genes are selectively constrained for their transcriptional noise, and that centrality therefore also influences the regulation of gene expression. Interestingly, it has been reported that central genes tend to be more duplicated (Vitkup et al. 2006). The authors proposed that such duplication events would have been favored as they would confer greater robustness to deleterious mutations in proteins. Our results are compatible with another, non exclusive, possible advantage: having more gene copies could reduce transcriptional noise by averaging the amount of transcripts produced by each gene copy (Raser and O’Shea 2005).

Network structure impacts transcriptional noise of constitutive genes

Whereas estimators of node centrality highlight gene-specific properties inside a given network, measures at the whole-network level enable the comparison of networks with distinct properties. We computed the size, diameter and transitivity for each annotated network in our data set (1,364 networks, Supplementary Material), as well as average measures of node scores (degree, hub score, authority score, closeness, betweenness) which we compare with the average F* measure of all constitutive nodes. The size of a network is defined as its total number of nodes, while diameter is the length of the shortest path between the two most distant nodes. Transitivity is a measure of connectivity, defined as the average of all nodes' clustering coefficients, defined for each node as the proportion of its neighbors that also connect to each other. Interestingly, while network size is positively correlated with average degree and transitivity (Kendall’s tau = 0.372, p-value < 2.2e-16 and Kendall's tau = 0.119, p-value = 2.807, respectively), diameter displays a positive correlation with average degree (Kendall's tau = 0.202, p-value < 2.2e-16) but a negative correlation with transitivity (Kendall's tau = -0.115, p-value = 2.237e-08). This is because diameter increases logarithmically with size, that is, addition of new nodes to large networks do not increase the diameter as much as additions to small networks. This suggests that larger networks are relatively more compact than smaller ones, and their constitutive nodes are therefore more connected. We find that average transcriptional noise correlates negatively with network size (Kendall's tau = -0.0594, p-value = 0.001376), while being independent of the diameter (Kendall's tau = 0.0125, p-value = 0.5366). Transcriptional noise is also strongly negatively correlated with all averaged centrality measures (**Table 3**). These results are in line with the node-based analyses, and show that the more connections a network has, the less stochastic the expression of the underlying genes is. This supports the view of Raser and Oshea (Raser and O'Shea 2005) that the gene-extrinsic, pathway-intrinsic level is functionally pertinent and needs to be distinguished from the globally extrinsic level.

We further asked whether genes with similar transcriptional noise tend to synthesize proteins that connect to each other (positive assortativity) in a given network, or on the contrary, tend to avoid each other (negative assortativity). We considered all Reactome pathways annotated to the mouse and estimated their respective F* assortativity. We found the mean assortativity to be significantly negative, with a value of -0.131 (one sample Wilcoxon rank test, p-value < 2.2e-16), meaning that proteins with different F* values tend to connect with each other (**Figure S3**). Maslov & Sneppen (Maslov and Sneppen 2002) reported a negative assortativity between hubs in protein-protein interaction networks, which they hypothesized to be the result of selection for reduced vulnerability to deleterious perturbations. In our data set, however, we find the assortativity of hub scores to be slightly but significantly positive (average of 0.060, one sample Wilcoxon rank test, p-value = 0.0002702, **Figure S3**), although with a large distribution of assortativity values. As we showed that hub scores correlates negatively with F* (**Table 2**), we asked whether the negative assortativity of hub proteins can at least partly explain the negative assortativity of F*. We found a significantly positive correlation between the two assortativity measures (Kendall's tau = 0.338, p-value < 2.2e-16). The relationship between the measures, however, is not linear. A Multivariate Adaptive Regression Spline was fitted to the two assortativity measures and resulted in a selected model with a strong positive correlation for hub score assortativity below -0.16, and virtually no correlation above (**Figure S3**), suggesting a distinct relationship between hub score and F* for negative and positive hub score assortativity. Negative assortativity of hub proteins contributes to a negative assortativity of SGE (Kendall's tau = 0.381, p-value < 2.2e-16), while for pathways with positive hub score assortativity the effect disappears (Kendall's tau = 0.052, p-value = 0.06282). While assortativity of F* is closer to 0 for pathways with positive assortativity of hub score, we note that it is still significantly negative (average = -0.047, one sample Wilcoxon test with p-value < 2.2e-16). This suggests the existence of additional constraints that act on the distribution of noisy proteins in a network.

Transcriptional noise is positively correlated with the evolutionary rate of proteins

In the yeast *Saccharomyces cerevisiae*, evolutionary divergence between orthologous coding sequences correlates negatively with fitness effect on knock-out strains of the corresponding genes (Hirsh and Fraser 2001), demonstrating that protein functional importance is reflected in the strength of purifying selection acting on it. Fraser et al (Fraser et al. 2004) studied transcription and translation rates of yeast genes and classified genes in distinct noise categories according to their expression strategies. They reported that genes with high fitness effect display lower expression noise than the rest. Following these pioneering observations, we hypothesized that genes under strong purifying selection at the protein sequence level should also be highly constrained for their expression and therefore display a lower transcriptional noise. To test this hypothesis, we correlated F* with the ratio of non-synonymous (Ka) to synonymous substitutions (Ks), as measured by sequence comparison between mouse genes and their human orthologs, after discarding genes with evidence for positive selection (n = 5). In agreement with our prediction, we report a significantly positive correlation between the Ka / Ks ratio and F* (**Figure 4**, Kendall's tau = 0.0619, p-value < 2.2e-16), that is, highly constrained genes display less transcriptional noise than fast evolving ones. This result demonstrates that genes encoding proteins under strong purifying selection are also more constrained on their transcriptional noise.

Older genes are less noisy

Evolution of new genes was long thought to occur via duplication and modification of existing genetic material (“evolutionary tinkering”, (Jacob 1977)). Evidence for *de novo* gene emergence is however becoming more and more common (Tautz and Domazet-Lošo 2011; Xie et al. 2012). *De novo* created genes undergo several optimization steps, including their integration into a regulatory network (Neme and Tautz 2013). We tested whether the historical process of incorporation of new genes into pathways impacts the evolution of transcriptional noise. We used the phylostratigraphic approach of Neme & Tautz (Neme and Tautz 2013), which categorizes genes into 20 strata, to compute gene age and tested for a correlation with F*. As older genes tend to be more conserved (Wolf et al. 2009), more central (according to the preferential attachment model of network growth

(Jeong et al. 2000; Jeong et al. 2001)) and more pleiotropic, we controlled for these confounding factors (**Figure 4**, Kendall's tau = -0.041, p-value = 1.406e-15 ; partial correlation controlling for Ka / Ks ratio, centrality measures and pleiotropy level). These results suggest that older genes are more deterministically expressed while younger genes are more noisy. While we cannot rule out that functional constraints not fully accounted for by the Ka / Ks ratio or unavailable functional annotations could explain at least partially the correlation of gene age and transcriptional noise, we hypothesise that the observed correlation result from ancient genes having acquired more complex regulation schemes through time. Such schemes include for instance negative feedback loops, which have been shown to stabilize gene expression and reduce expression noise (Becskei and Serrano 2000; Thattai and Oudenaarden 2001).

Position in the protein network is the main driver of transcriptional noise

In order to jointly assess the effect of network topology, epigenomic factors, Ka / Ks ratio and gene age, we modeled the patterns of transcriptional noise as a function of multiple predictive factors within the linear model framework. In order to avoid overfitting due to a large number of explanatory variables, and because some of these variables are intrinsically correlated and can lead to colinearity issues, we performed data reduction procedures prior to modeling. For network variables, we used as synthetic measures of node centrality the first four principal components of the principal component analysis (PCA), explaining together 93% of the total inertia (**Figure S2**). As transcription factors and histone marks data are binary (presence / absence for each gene), we performed a logistic PCA for both type of variables (Landgraf and Lee 2015). For transcription factors, we selected the three first components, which explained 78% of deviance (**Figure S3**). The loads on the first component (PC1) are all negative, meaning that PC1 captures a global correlation trend and does not discriminate between TFs. The second component PC2 is dominated by TCFC (positive loading) and Oct4 (negative loading), while the third component PC3 is dominated by Esrrb (positive loading) and MYC, nMyc and E2F1 (negative loadings). For histone marks, the two first components explained 95% of variance and were therefore retained (**Figure S4**). PC1 is dominated by marks H3K27me3 and H3K9me3 linked to gene repression (negative loadings) and PC2 by marks H3K4me1 and H3K4me3 linked to gene activation (positive loadings).

We fitted a linear model with F* as a response variable, Ka / Ks ratio, gene age, the four synthetic network centrality measures, the three synthetic variables capturing the transcription factor binding evidences and the two synthetic variables capturing the presence of histone marks as explanatory variables. We also included the mean gene expression in order to account for spurious correlation of F* with mean expression. We find that despite the intrinsic accounting of F* for mean expression, there is still a significant positive correlation with mean gene expression, which was not detected by Kendall's rank correlation test (see above). The corresponding coefficient, however, is very low (0.0003, **Table 4**). In agreement with our single variable analyses, we report that Ka / Ks ratio and gene age are significantly positively and negatively correlated with transcriptional noise, respectively (**Table 4**). We further find that the first component of the network PCA analysis has a significant positive effect on F*. This measure essentially captures the effect of node degree, hub and transitivity scores (**Figure S2**); this result is therefore also consistent with single variable analyses. The second component of the logistic PCA of transcription factor binding evidence, as well as the first component of the logistic PCA on histone marks are also found to be significant (**Table 4**), which confirms the effect of these variables when other factors are accounted for. The coefficient associated with transcription factor PC2 is positive, which indicates that TFs increase transcriptional noise, in particular TCFC which has the highest loading on PC2. The coefficient associated with histone marks PC1, however is negative. Yet the largest loadings of the variables on this component are negative (H3K27me3 and H3K9me3), implying that these histones marks are associated with a higher transcriptional noise, as found by individual tests.

Altogether, the linear model with all variables explained 3.93% of the total variance. This small value indicates either that gene idiosyncrasies largely predominate over general effects, or that our estimates of transcriptional noise have a large measurement error, or both. An analysis of variance shows that the centrality variable explains the largest part of the variance (1.66% variance explained for the first synthetic variable, Fisher’s test p-value = 9.552e-15 and 0.11% for the second synthetic variable, p-value = 0.0410). Mean gene expression only explained 0.11% of the total variance (Fisher’s test p-value = 0.0386). Gene age only explains 0.31% of the variance (Fisher’s test p-value = 1.432e-09) and functional constraints 1% (Ka / Ks variable, Fisher’s test p-value = 0.0007). Transcription factors explain 0.19% of variance (Fisher’s test p-value = 0.0079) and histone marks 0.48% (Fisher’s test p-value = 2.665e-5). This suggests that, among all factors tested, position in protein network is the main driver of the evolution of gene-specific stochastic expression.

We further included the effect of three-dimensional organization of the genome in order to assess whether it could be a confounding factor. We developed a correlation model allowing for genes in contact to have correlated values of transcriptional noise. The correlation model was fitted together with the previous linear model in the generalized least square (GLS) framework. This model allows for one additional parameter, λ, which captures the strength of correlation due to three-dimensional organization of the genome (see Methods). The estimate of λ was found to be 0.0036, which means that the spatial autocorrelation of transcriptional noise is low on average. This estimate is significantly higher than zero, and model comparison using Akaike’s information criterion favors the linear model with three-dimensional correlation, yet with very low support (AIC = 6403.452 vs.

AIC = 6403.859 for a linear model without three-dimensional correlation). Consistently, accounting for this correlation does not change significantly our estimates (**Table 4**), confirming network centrality measures as the main factor explaining the distribution of transcriptional noise.

Analysis of bone marrow-derived dendritic cells supports the generality of the results

We assessed the reproducibility of our results by analyzing an additional single-cell transcriptomics data set of 95 unstimulated bone marrow-derived dendritic cells (BMDC) (Shalek et al. 2014). After filtering (see Methods), the data set consisted of 11,640 genes. Using the same normalization procedure as for the Sasagawa data set, we nonetheless report a weak but significant negative correlation between F* and the mean expression (-0.068, p-value < 2.2e-16). We fitted a generalized linear model as for the embryonic stem cell (ESC) data set, with the exception that no epigenomic data was available for this cell type. Results of this model are very similar to the ones with the ESC data set: the model explains 3.24% of the variance, with 1.42% explained by network measures, and all effects are similar in direction and intensity (**Table S1**). When taking 3D genome correlations into account, we estimated a low correlation coefficient as for the ESC dataset (lambda = 0.0025), and the AIC favored the model without correlation. The mean gene expression is not found to be significant when taken together with other parameters in the BMDC data set. Interestingly, we find that the second and fourth principal components of the network analysis are also significant with this data set. We note that values of the “closeness” variable, which are for this dataset positively correlated with “betweenness” values, while they are negatively correlated for the ESC dataset. While these results support the generality of our observations, they also illustrate that in details, the structure of translational noise may vary in a cell type-specific manner.

Biological, not technical noise is responsible for the observed patterns

The variance in gene expression measured from single-cell transcriptomics is a combination of biological and technical variance. While the two sources of variance are a priori independent, gene-specific technical variance has been observed in micro-array experiments (Pozhitkov et al. 2007) making a correlation of the two types of variance plausible. If similar effects also affect RNA-Seq experiments, technical variance could be correlated to gene function and therefore act as a covariate in our analyses. In order to assess whether this is the case, we used the dataset of Shalek et al (Shalek et al. 2013), which contains both single-cell transcriptomics and 3 replicates of 10,000 pooled-cell RNA sequencing. In traditional RNA sequencing, which is typically performed on pooled populations of several thousands of cells, biological variance is averaged out so that the resulting measured variance between replicates is essentially the result of technical noise. We computed the mean and variance in expression of each gene across the three populations of cells. By plotting the variance versus the mean in log-space, we were able to compute a “technical” F* (F_t^*) value for each gene (Methods). We fitted linear models with and without 3D genome correlation as for the single cell data, using F_t^* instead of F*. We report that no variable but the mean gene expression had a significant effect on F_t^* , yet with a very low size effect (**Table S2**). In addition, there was no enrichment of the 10th and 90th F_t^* percentiles for any particular pathway or GO term. These results therefore support our conclusion that the correlations we observe are due to variations that are biological, not technical.

Discussion

Throughout this work, we provided the first genome-wide evolutionary and systemic study of transcriptional noise, using mouse cells as a model. We have shown that transcriptional noise correlates with functional constraints both at the level of the gene itself via the protein it encodes, but also at the level of the pathway(s) the gene belongs to. We further discuss here potential confounding factors in our analyses and argue that our results are compatible with selection acting to reduce noise-propagation at the network level.

In this study, we exhibited several factors explaining the variation in transcriptional noise between genes. While highly significant, the effects we report are of small size, and a complex model accounting for all tested sources of variation only explains a few percent of the total observed variance. There are several possible explanations for this reduced explanatory power: (1) transcriptional noise is a proxy for noise in gene expression, at which selection occurs (**Figure 1**). As transcriptional noise is not randomly distributed across the genome, it must constitute a significant component of expression noise, in agreement with previous observations (Blake et al. 2003; Newman et al. 2006). Translational noise, however, might constitute an important part of the expression noise and was not assessed in this study. (2) Gene expression levels were assessed on embryonic stem cells in culture. Such an experimental system may result in gene expression that differs from that in natural conditions under which natural selection acted. (3) Functional annotations, in particular pathways and gene interactions are incomplete, and network-based measures have most likely large false positive and negative error rates. (4) While the newly introduced F* measure allowed us to assess the distribution of transcriptional noise independently of the average mean expression – therefore constituting an improvement over previous studies – it does not capture the full complexity of SGE. Explicit modeling, for instance based in the Beta-

Poisson model (Vu et al. 2016) is a promising avenue for the development of more sophisticated quantitative measures.

In a pioneering study, Fraser et al, followed by Shalek et al, demonstrated that essential genes whose deletion is deleterious, and genes encoding subunits of molecular complexes (Fraser et al. 2004) as well as housekeeping genes (Shalek et al. 2013) display reduced gene expression noise. Our findings go beyond these early observations by providing a statistical assessment of the joint effect of multiple explanatory factors. Our analyses reveal that network centrality measures are the explanatory factors that explained the most significant part of the distribution of transcriptional noise in the genome. This suggests that selection at the pathway level is a widespread phenomenon that drives the evolution of SGE at the gene level. This multi-level selection mechanism, we propose, can be explained by selection against noise propagation within networks. It has been experimentally demonstrated that expression noise can be transmitted from one gene to another gene with which it is interacting (Pedraza and van Oudenaarden 2005). Large noise at the network level is deleterious (Barkai and Leibler 1999) but each gene does not contribute equally to it, thus the strength of selective pressure against noise varies among genes in a given network. We have shown that highly connected, “central” proteins typically display reduced transcriptional noise. Such nodes are likely to constitute key players in the flow of noise in intra-cellular networks as they are more likely to transmit noise to other components. In accordance with this hypothesis, we find genes with the lowest amount of transcriptional noise to be enriched for top-level functions, in particular involved in the regulation of other genes.

These results have several implications for the evolution of gene networks. First, this means that new connections in a network can potentially be deleterious if they link genes with highly stochastic expression. Second, distinct selective pressures at the “regulome” and “interactome” levels (**Figure 1**) might act in opposite direction. We expect genes encoding highly connected proteins to have more complex regulation schemes, in particular if their proteins are involved in several biological pathways. In accordance, several studies demonstrated that expression noise of a gene positively correlates with the number of transcription factors controlling its regulation (Sharon et al. 2014), a correlation that we also find significant in the data set analysed in this work. Central genes, while being under negative selection against stochastic behavior, are then more likely to be controlled by numerous transcription factors which increase transcriptional noise. As a consequence, if the number of connections at the interactome level is correlated with the number of connections at the regulome level, we predict the existence of a trade-off in the number of connections a gene can make in a network. Alternatively, highly connected genes might evolve regulatory mechanisms allowing them to uncouple these two levels: negative feedback loops, for instance, where the product of a gene down-regulates its own production have been shown to stabilize expression and significantly reduce stochasticity (Becskei and Serrano 2000; Dublanche et al. 2006; Tao et al. 2007). We therefore predict that negative feedback loops are more likely to occur at genes that are more central in protein networks, as they will confer greater resilience against high SGE, which is advantageous for this class of genes.

Our results enabled the identification of possible selective pressures acting on the level of stochasticity in gene expression. The mechanisms by which the amount of stochasticity can be controlled remain however to be elucidated. We evoked the existence of negative feedback loops which reduce stochasticity and the multiplicity of upstream regulator which increase it. Recent work by Wolf et al (Wolf et al. 2015) and Metzger et al (Metzger et al. 2015) add further perspective to this scheme. Wolf and colleagues found that in *Escherichia coli* noise is higher for natural than experimentally evolved promoters selected for their mean expression level. They hypothesized that higher noise is selectively advantageous in case of changing environments. On the other hand, Metzger and colleagues performed mutagenesis experiments and found signature of selection for reduced noise in natural populations of *Saccharomyces cerevisae*. These seemingly opposing results combined with our observations provide additional evidence that the amount of stochasticity in the expression of single genes has an optimum, as high values are deleterious because of noise propagation in the network, whilst lower values, which result in reduced phenotypic plasticity, are suboptimal in case of changing environment.

Conclusion

Using a new measure of transcriptional noise, our results demonstrate that the position of the protein in the interactome is a major driver of selection against stochastic gene expression. As such, transcriptional noise is an essential component of the phenotype, in addition to the mean expression level and the actual sequence and structure of the encoded proteins. This is currently an under-appreciated phenomenon, and gene expression studies that focus only on the mean expression of genes may be missing key information about expression diversity. The study of gene expression must consider changes in noise in addition to change in mean expression level as a putative explanation for adaptation. Further work aiming to unravel the exact structure of the regulome is however needed in order to fully understand how transcriptional noise is generated or inhibited.

Material and Methods

Single-cell gene expression data set

We used the dataset generated by Sasagawa et al. (Sasagawa et al. 2013) retrieved from the Gene Expression Omnibus repository (accession number GSE42268). We analyzed expression data corresponding to embryonic stem cells in G1 phase, for which more individual cells were sequenced. A total of 17,063 genes had non-zero expression in at least one of the 20 single cells. Similar to Shalek et al (Shalek et al. 2014), a filtering procedure was performed where only genes whose expression level satisfied log(FPKM+1) > 1.5 in at least one single cell were kept for further analyses. This filtering step resulted in a total of 13,660 appreciably expressed genes for which transcriptional noise was evaluated.

Measure of transcriptional noise

The expression mean (μ) and variance (σ^2) of each gene over all single cells were computed. A linear model was fitted on the log-transformed means and variances in order to estimate the coefficients of the power law regression:

$$\sigma^2 = a.\mu^b \quad \text{(eqn 1)}$$

$$\log(\sigma^2) = \log(a) + b.\log(\mu) \quad \text{(eqn 2)}$$

We defined F* as the ratio of the observed variance and the predicted variance:

$$F^* = \frac{\sigma^2}{a.\mu^b} \quad \text{(eqn 3)}$$

F* can be seen as a general expression for the Fano factor (a = b = 1) and noise measure (a = 1, b = 2). F* is the stochasticity measure unit with which we produced our results, after estimating the a and b parameters from the data.

Genome architecture

The mouse proteome from Ensembl (genome version: mm9) was used in order to get coordinates of all genes. The Hi-C dataset for embryonic stem cells (ES) from Dixon et al (Dixon et al. 2012) was used to get three-dimensional domain information. Two genes were considered in proximity in one dimension (1D) if they are on the same chromosome and no protein-coding gene was found between them. The primary distance (in number of nucleotides) between their midpoint coordinates was also recorded as 1D a distance measure between the genes. Two genes were considered in proximity in three dimensions (3D) if the normalized contact number between the two windows the genes belong was non-null. Two genes belonging to the same window were considered in proximity. We further computed the relative difference of stochastic gene expression between two genes by computing the ratio $(F_2^* - F_1^*)/(F_2^* + F_1^*)$. For each chromosome, we independently tested if there was a correlation between the primary distance and the relative difference in stochastic gene expression with a Mantel test, as implemented in the ade4 package (Dray and Dufour, 2007). In order to test whether genes in proximity (1D and 3D) had more similar transcriptional noise than distant genes, we contrasted the relative differences in transcription noise between pairs of genes in proximity and pairs of distant genes. As we test all pairs of genes, we performed a randomization procedure in order to assess the significance of the observed differences by permuting the rows and columns in the proximity matrices 1,000 times. Linear models accounting for spatial interactions with genes were fitted using the generalized least squares (GLS) procedure as implemented in the "nlme" package for R (Pinheiro et al 2016). A correlation matrix between all tested genes was defined as $G = \{g_{i,j}\}$, where $g_{i,j}$ is the correlation between genes i and j. We defined $g_{i,j} = 1 - \exp(-\lambda \delta_{i,j})$, where $\delta_{i,j}$ takes 1 if genes i and j are in proximity, 0 otherwise. Parameter λ was estimated jointly with other model parameters, it measures the strength of the genome "spatial" correlation. Parameters were estimated using the maximum likelihood (ML) procedure, instead of the default restricted maximum likelihood (REML) in order to perform model comparison using Akaike's information criterion (AIC).

Transcription factors and histone marks

Transcription factor (TF) mapping data from the Ensembl regulatory build (Zerbino et al. 2015) were obtained via the biomaRt package for R. We used the Grch37 build as it contained data for stem cells epigenomes. Genes were considered to be associated with a given TF when at least one binding evidence was present in the 3 kb upstream flanking region. Transcription factors associated with less than 5 genes for which transcriptional noise could be computed were not considered further. A similar mapping was performed for histone marks by counting the evidence of histone modification in the 3 kb upstream and downstream regions of each gene. A logistic principal component analysis was conducted on the resulting binary contingency tables using the logisticPCA package for R (Landgraf and Lee 2015), for TF and histone marks separately. Principal components were used to define synthetic variables for further analyses.

Biological pathways and network topology

The 13,660 Ensembl ids in our dataset were mapped to 13,136 Entrez ids. We kept only genes with unambiguous mapping, resulting in 11,032 Entrez ids for the Reactome pathway analysis. We defined genes either in the top 10% least noisy or in the top 10% most noisy as candidate sets and used the Reactome PA package (Yu and He 2015) to search the mouse Reactome database for overrepresented pathways with a 1% false discovery rate.

Thirteen thousand six hundred and sixty Ensembl ids mapped to a total of 29,859 UniProt ids. For network analyses, we removed UniProt ids which were not annotated to the Reactome database, resulting in a total of 4,929 UniProt ids after this first step. We then removed genes that mapped ambiguously from Ensembl to UniProt, retaining 3,959 Ensembl / UniProt ids for which we computed centrality measures. At the network level, size, transitivity and diameter could be calculated for every pathway using a combination of three R packages ("pathview" (Luo 2013), "igraph" (Csardi 2015**)** and "graphite" (Sales et al 2016)). As the calculation of assortativity does not handle missing data (that is, nodes of the pathway for which no value could be computed), we computed assortativity on the sub-network with nodes for which data were available. A principal component analysis was conducted on all network centrality measures using the ade4 package for R (Dray and Dufour 2007). Models of F* assortativity measures were fitted and compared using Multivariate Adaptive Regression Splines, as implemented in the "earth" package in R (Milborrow 2016).

Gene Ontology Enrichment

Eight thousand three hundreds and twenty five out of the 13,660 genes were associated with Gene Ontology (GO) terms. We tested genes for GO terms enrichment at both ends of the F* spectrum using the same threshold percentile of 10% low / high noise genes as we did for the Reactome analysis. We carried out GO enrichment analyses using two different algorithms: "Parent-child" (Grossmann et al. 2007) and "Weight01", a mixture of two algorithms developed by Alexa et al (Alexa et al. 2006). We kept only the terms that appeared simultaneously on both Parent-child and Weight01 under 10% significance level, controlling for multiple testing using the FDR method (Benjamini and Hochberg 1995).

Sequence divergence

The Ensembl's Biomart interface was used to retrieve the proportion of non-synonymous (Ka) and synonymous (Ks) divergence estimates for each mouse gene relative to the human ortholog. This information was available for 13,136 genes.

Gene Age

The relative taxonomic ages of the mouse genes have been computed and is available in the form of 20 Phylostrata (Neme and Tautz 2013). Each Phylostratum corresponds to a node in the phylogenetic tree of life. Phylostratum 1 corresponds to "All cellular organisms" whereas

Phylostratum 20 corresponds to "*Mus musculus*", with other levels in between. We used this published information to assign each of our genes to a specific Phylostratum and used this as a relative measure of gene age: Age = 21 - Phylostratum, so that an age of 1 corresponds to genes specific to *M. musculus* and genes with an age of 20 are found in all cellular organisms.

Linear modeling

We simultaneously assessed the effect of different factors on transcriptional noise by fitting linear models to the gene-specific F* estimates. To avoid colinearity issues of intrinsically correlated explanatory variables, we used a principal component regression approach, using principal components analysis to reduce the number of input variables. We built a linear model with F* as a response variable and the four first components of network centrality measures, three first components of TF binding variables, two first components of histone marks variables, as well as the Ka / Ks ratio and gene age. As the fitted model displayed significant departure to normality, it was further transformed using the Box-Cox procedure ("boxcox" function from the MASS package for R (Venables and Ripley 2002)). Residues of the selected model had independent residue distributions (Ljung-Box test, p-value = 0.1008) but still displayed significant departure to normality (Shapiro-Wilk test, p-value = 1.751e-5), and heteroscedasticity (Harrison-McCabe test, p-value = 0.00067). In order to assess whether these departures from the Gauss-Markov assumptions could bias our results, we used two complementary approaches. First we used the "robcov" function of the "rms" package in order to get robust estimates of the effect significativity (Harrel 2016). Second, we performed a quantile regression using the "rq" function (parameter tau set to 0.5, equivalent to a median regression) of the "quantreg" package for R (Koenker, 2016). As quantile regression results were systematically consistent with linear regression analyses, we only report results from the latter.

Additional data sets

The aforementioned analyses were additionally conducted on the data set of Shalek et al (Shalek et al. 2014). Following the filtering procedure established by the authors in the original paper, genes which did not satisfied the condition of being expressed by an amount such that log(TPM+1) > 1 in at least one of the 95 single cells were further discarded, where TPM stands for transcripts per million. This cut-off threshold resulted in 11,640 genes being kept for investigation. The rest of the analyses was conducted in the same way as in Sasagawa's data set.

Data and program availability

All datasets and scripts to reproduce the results of this study are available at Figshare, under the DOI 10.6084/m9.figshare.4587169.

Authors contributions

GVB and JYD designed the experiments and wrote the manuscript. GVB, NP and JYD conducted the analyses.

Acknowledgements

The authors would like to thank Rafiq Neme-Garrido, Frederic Bartels and Estelle Renaud for fruitful discussions about this work, Andrew Landgraf for help with the logistic PCA analysis as well as Diethard Tautz for comments on an earlier version of this manuscript. JYD acknowledges funding from the Max Planck Society. This work was supported by the German Research Foundation (DFG), within the priority program (SPP) 1590.

References

Arkin A, Ross J, Mcadams HH. 1998. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage L-Infected Escherichia coli Cells. Genetics 149:1633–1648.

Barabási A-L, Albert R. 1999. Emergence of Scaling in Random Networks. Science 286:509–513.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5:101–113.

Bar-even A, Paulsson J, Maheshri N, Carmi M, Shea EO, Pilpel Y, Barkai N. 2006. Noise in protein expression scales with natural protein abundance. Nature genetics 38:636–643.

Barkai N, Leibler S. 1999. Circadian clocks limited by noise. Nature 403:267–268.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.

Becskei A, Kaufmann BB, van Oudenaarden A. 2005. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nature Genetics 37:937–944.

Becskei A, Serrano L. 2000. Engineering stability in gene networks by autoregulation. Nature 405:590–593.

Blake WJ, Kærn M, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. Nature 422:633–637.

Chubb JR, Trcek T, Shenoy SM, Singer RH. 2006. Transcriptional Pulsing of a Developmental Gene. Current Biology 16:1018–1025.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376–380.

Dublanche Y, Michalodimitrakis K, Kümmerer N, Foglierini M, Serrano L. 2006. Noise in transcription negative feedback loops: simulation and experimental analysis. Molecular systems biology 2:41–41.

Eldar A, Elowitz MB. 2010. Functional roles for noise in genetic circuits. Nature 467:167–173.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic Gene Expression in a Single Cell. Science 297:1183–1186.

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise Minimization in Eukaryotic Gene Expression. PLoS Biology 2:0834–0838.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary Rate in the Protein Interaction Network. Science 296:750–752.

Gillesple DT. 1977. Exact Simulation of Coupled Chemical Reactions. The Journal of Physical Chemistry 81:2340–2361.

Guimera R, Amaral LAN. 2005. Functional cartography of complex metabolic networks. Nature 433:895–900.

Hahn MW, Conant GC, Wagner A. 2004. Molecular Evolution in Large Genetic Networks: Does Connectivity Equal Constraint? Journal of Molecular Evolution 58:203–211.

Hahn MW, Kern AD. 2004. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. Molecular Biology and Evolution 22:7–10.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. Nature 402:C47–C52.

Hebenstreit D. 2013. Are gene loops the cause of transcriptional noise? Trends in Genetics 29:333–338.

Hirsh A, Fraser H. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.

Jacob F. 1977. Evolution and Tinkering. Science 196:1161–1166.

Jeong H, Mason SP, Barabási a L, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. 2000. The large-scale organization of metabolic networks. Nature 407:651–654.

Jovelin R, Phillips PC. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. Genome biology 10:R35–R35.

Joy MP, Brock A, Ingber DE, Huang S. 2005. High-betweenness proteins in the yeast protein interaction network. Journal of Biomedicine and Biotechnology 2005:96–103.

Kaufmann BB, van Oudenaarden A. 2007. Stochastic gene expression: from single molecules to the proteome. Current opinion in genetics & development 17:107–112.

Kepler TB, Elston TC. 2001. Stochasticity in Transcriptional Regulation : Origins, Consequences, and Mathematical Representations. Biophysical Journal 81:3116–3136.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2013. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. Science 603:1938–1941.

Landgraf AJ, Lee Y. 2015. Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. arXiv:1510.06112 [stat] [Internet]. Available from: http://arxiv.org/abs/1510.06112

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Molecular systems biology 4:170–170.

Maslov S, Sneppen K. 2002. Specificity and Stability in Topology of Protein Networks. Science 296:910–913.

Mcadams HH, Arkin A. 1997. Stochastic mechanisms in gene expression. Proceedings of the National Academy of Sciences of the United States of America 94:814–819.

Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. Nature 521:344–347.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC genomics 14:117–117.

Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, Derisi JL, Weissman JS. 2006. Single-cell proteomic analysis of S . cerevisiae reveals the architecture of biological noise. Nature 441:840–846.

Norman TM, Lord ND, Paulsson J, Losick R. 2015. Stochastic Switching of Cell Fate in Microbes. Annual review of microbiology 69:381–403.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, Oudenaarden AV. 2002. Regulation of noise in the expression of a single gene. Nature genetics 31:69–73.

Pál C, Papp B, Hurst LD. 2001. Highly Expressed Genes in Yeast Evolve Slowly. Genetics 158:927–931.

Pedraza JM, van Oudenaarden A. 2005. Noise propagation in gene networks. Science 307:1965–1969.

Pombo A, Dillon N. 2015. Three-dimensional genome architecture: players and mechanisms. Nature Reviews Molecular Cell Biology 16:245–257.

Pozhitkov, Alex E., Tautz D, Noble, Peter A. 2007. Oligonucleotide microarrays: widely appliedçpoorly understood. B RIEFINGS IN FUNC TIONAL GENOMICS AND P ROTEOMICS . 6:141–148.

Raj A, Oudenaarden AV. 2008. Review Nature , Nurture , or Chance : Stochastic Gene Expression and Its Consequences. Cell 135:216–226.

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. 2006. Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biology 4:e309–e309.

Raser JM, O'Shea EK. 2005. Noise in Gene Expression: Origins, Consequences, and Control. Science 309.

Sánchez A, Kondev J. 2008. Transcriptional control of noise in gene expression. Proceedings of the National Academy of Sciences of the United States of America 105:5081–5086.

Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. 2013. Quartz-Seq : a highly reproducible and sensitive single-cell RNA sequencing method , reveals non- genetic gene-expression heterogeneity. Genome Biology 14:R31–R31.

Sauer U, Heineman M, Zamboni N. 2007. Getting Closer to the Whole Picture. Science 316:550–551.

Shahrezaei V, Swain PS. 2008. The stochastic nature of biochemical networks. Curr. Opin. Biotechnol. 19:369–374.

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498:236–240.

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, et al. 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510:363–369.

Sharon E, Van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Research 24:1698–1706.

So L, Ghosh A, Zong C, Sepúlveda LA, Segev R, Golding I. 2011. General properties of transcriptional time series in Escherichia coli. Nature Genetics 43:554–560.

Spudich JL, Koshland DEJ. 1976. Non-genetic individuality: chance in the single cell. Nature:467–471.

Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. 2011. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. Science 332:472–474.

Taniguchi Y, Choi PJ, Li G, Chen H, Babu M, Hearn J, Emili A, Xie XS. 2011. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science (New York, N.Y.) 329:533–539.

Tao Y, Zheng X, Sun Y. 2007. Effect of feedback regulation on stochastic gene expression. J. Theor. Biol. 247:827–836.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nature reviews. Genetics 12:692–702.

Thattai M, Oudenaarden AV. 2001. Intrinsic noise in gene regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 98:8614–8619.

Thattai M, Oudenaarden AV. 2004. Stochastic Gene Expression in Fluctuating Environments. Genetics 167:523–530.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome biology 7:R39–R39.

Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y. 2016. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics:1–8.

Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proceedings of the National Academy of Sciences 108:E67–E76.

Wolf L, Silander OK, van Nimwegen EJ. 2015. Expression noise facilitates the evolution of gene regulation. eLife 4:1–48.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proceedings of the National Academy of Sciences of the United States of America 106:7273–7280.

Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li CY. 2012. Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. PLoS Genetics 8:e1002942.-e1002942.

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. 2007. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Computational Biology 3:713–720.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. Genome Biol. 16:56.

Tables

Table 1: GO terms significantly enriched in the 10% genes with lowest transcriptional noise.

Ontology	GO ID	GO term	FDR Fisher "parent-child"	FDR Fisher "weight01"
MF	GO:0003676	nucleic acid binding	2.406E-03	1.475E-08
MF	GO:0003735	structural constituent of ribosome	6.099E-03	1.708E-05
BP	GO:0006334	nucleosome assembly	3.816E-03	1.380E-02
BP	GO:0002227	innate immune response in mucosa	6.727E-03	2.018E-02
BP	GO:0006412	translation	1.257E-02	1.380E-02
CC	GO:0000788	nuclear nucleosome	3.493E-05	2.587E-05

Note: FDR: False Discovery Rate. MF: Molecular Function. BP: Biological Process. CC: Cellular Compartment.

Table 2: Correlation of transcriptional noise with genes centrality measures and pleiotropy.

Measure	Correlation with F*	p-value
Degree	-0.069	3.192E-10
Hub score	-0.068	6.132E-10
Authority score	-0.064	6.151E-09
Closeness	-0.004	7.305E-01
Betweenness	-0.017	1.303E-01
Pleiotropy	-0.046	5.069E-05

Note: All correlations are computed using Kendall’s rank correlation test.

Table 3: Correlation of average transcriptional noise with pathway centrality measures.

Measure	**Correlation with average F***	**p-value**
Size	-0.059	1.376E-03
Diameter	0.012	5.366E-01
Average degree	-0.172	8.944E-21
Average hub score	-0.188	1.724E-24
Average authority score	-0.166	2.487E-19
Average closeness	0.050	6.500E-03
Average betweenness	-0.166	2.487E-19
Average pleiotropy	-0.137	1.276E-13

Note: All correlations are computed using Kendall's rank correlation test.

Table 4: Linear models of transcriptional noise with genomic and epigenomic factors.

	OLS			GLS			Variance
	Coefficient	S.E.	p-value	Coefficient	S.E.	p-value	
(Intercept)	0.5079	0.1130	<0.0001	0.5128	0.1077	<0.0001	
Mean expression	0.0003	0.0001	0.0002	0.0003	0.0001	0.0001	0.12%
Network PC1	0.0485	0.0066	<0.0001	-0.0482	0.0066	<0.0001	1.66%
Network PC2	-0.0141	0.0103	0.1724	-0.0141	0.0106	0.1867	0.11%
Network PC3	0.0036	0.0096	0.7066	0.0034	0.0099	0.7340	0.00%
Network PC4	-0.0065	0.0104	0.531	-0.0073	0.0108	0.5025	0.02%
TF PC1	0.0029	0.0038	0.4524	0.0029	0.0035	0.4152	0.00%
TF PC2	0.0064	0.0028	0.0206	0.0064	0.0027	0.0169	0.19%
TF PC3	0.0009	0.0038	0.8155	0.0007	0.0037	0.8406	0.02%
Histone PC1	-0.0034	0.0009	0.0001	-0.0034	0.0009	0.0001	0.48%
Histone PC2	0.0003	0.0015	0.8325	0.0004	0.0012	0.7693	0.01%
Ka / Ks	-0.0209	0.0048	<0.0001	0.3683	0.1083	0.0007	1.00%
Age	0.3665	0.1027	0.0004	-0.0211	0.0046	<0.0001	0.31%

Note: OLS: Ordinary Least Squares. GLS: Generalized Least Squares. Network PC1-4: principal components of the principal component analysis (PCA) on network measures. TF PCA1-3: principal components of the logistic PCA on transcription factors binding evidences. Histone PC1-2: principal components of the logistic PCA on histone modification marks. S.E.: standard error.

Figures

Figure 1: A systemic view of gene expression.

Figure 2: Transcriptional noise and mean gene expression. A) Measures of noise plotted against the mean gene expression for each gene, in logarithmic scales together with corresponding regression lines: variance, Fano factor (variance / mean), noise (square of the coefficient of variation, variance / mean^2) and F* (this study). B) Distribution of F* over all genes in this study. Vertical line corresponds to F* = 1.

Figure 3: Enriched pathways in the 10% genes with lowest transcriptional noise.

Figure 4: Correlation of F* with significant principal components of network centrality measures, transcription factors binding evidences and histone marks presence, as well as gene age and Ka / Ks ratio.

Supplementary material:

Table S1: Linear models of transcriptional noise with genomic factors for the bone marrow-derived dendritic cells data set. Legend as in Table 4.

Table S2: Linear models of transcriptional noise with genomic factors with pooled RNA-Seq data. Legend as in Table 4.

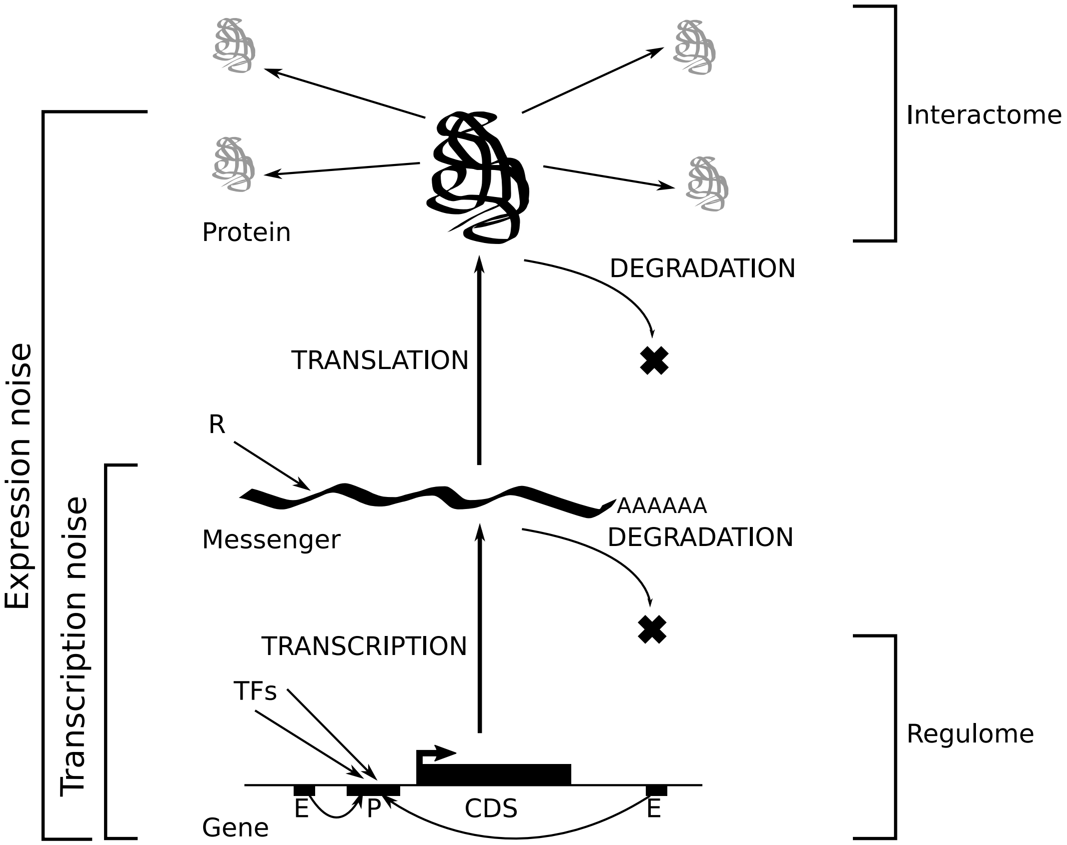
Figure S1: Impact of genome organization on the distribution of transcriptional noise. The x-axis shows the mean relative difference in transcriptional noise. Vertical lines show observed values and histograms the distribution over 1,000 permutations (see Methods). Left panel: distribution for neighbor genes along the genome. Right panel: distribution for genes in contact in three-dimensions.

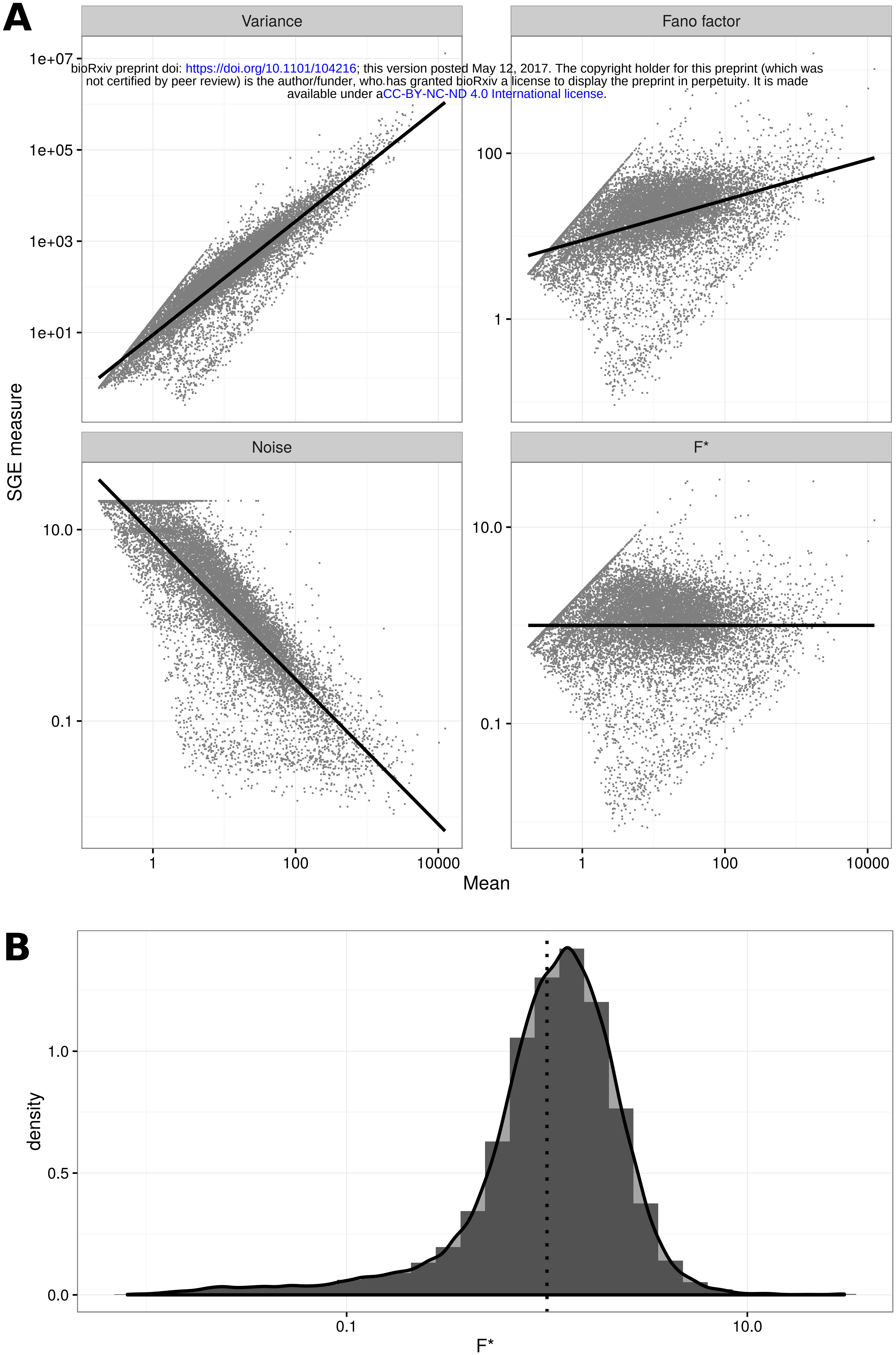
Figure S2: Principal component analysis of network measures. A) Proportion of deviance explained by models with 1, 2, etc. principal components. B) Loadings of each variable on the 2 first components. C) Loadings of each variable on the 2nd and 3rd principal components.

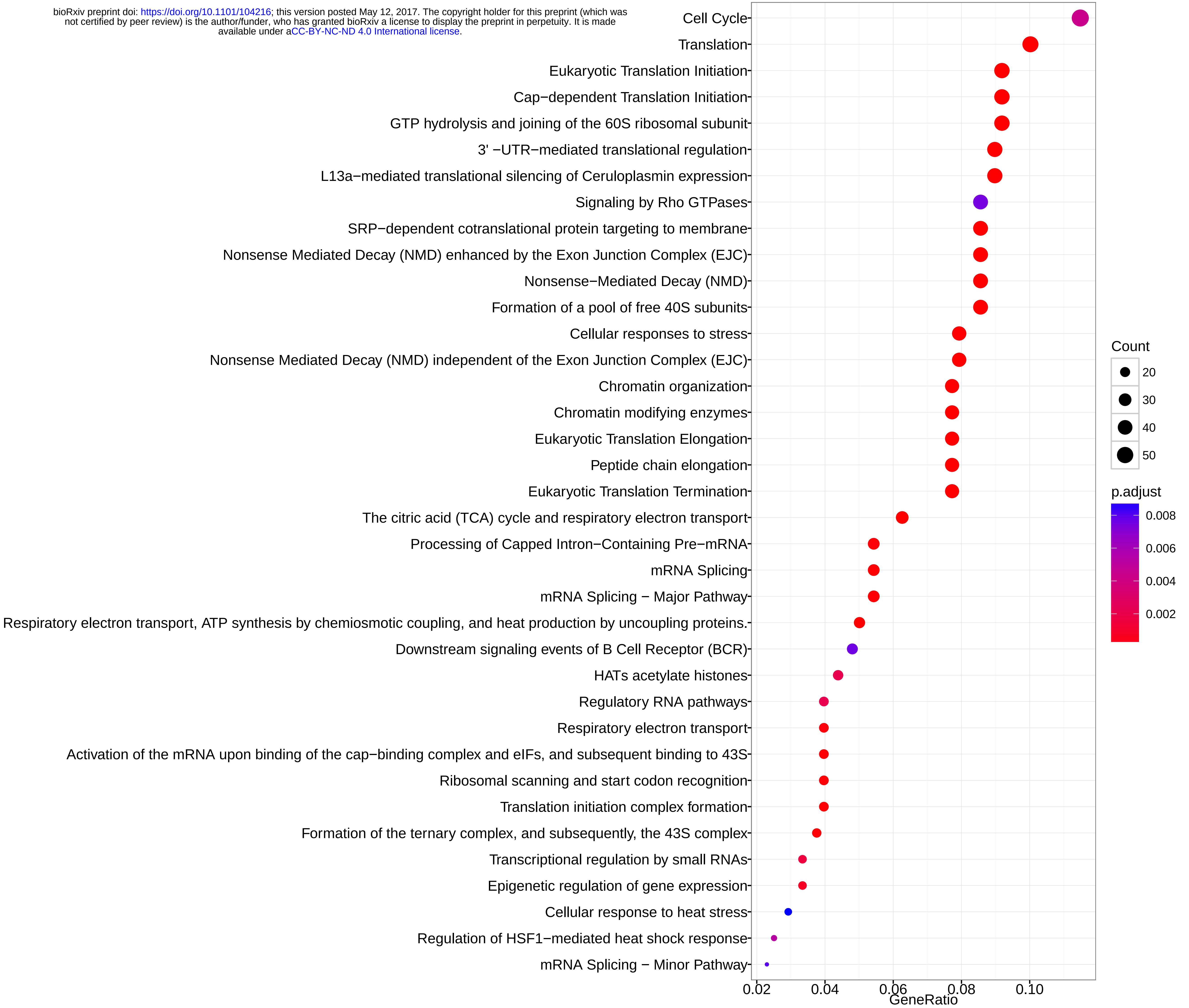
Figure S3: Logistic principal component analysis of transcription factor binding evidences. A) Proportion of deviance explained by models with 1, 2, etc. principal components. B) Loadings of each variable on the 2 first components. C) Loadings of each variable on the 2nd and 3rd principal components.

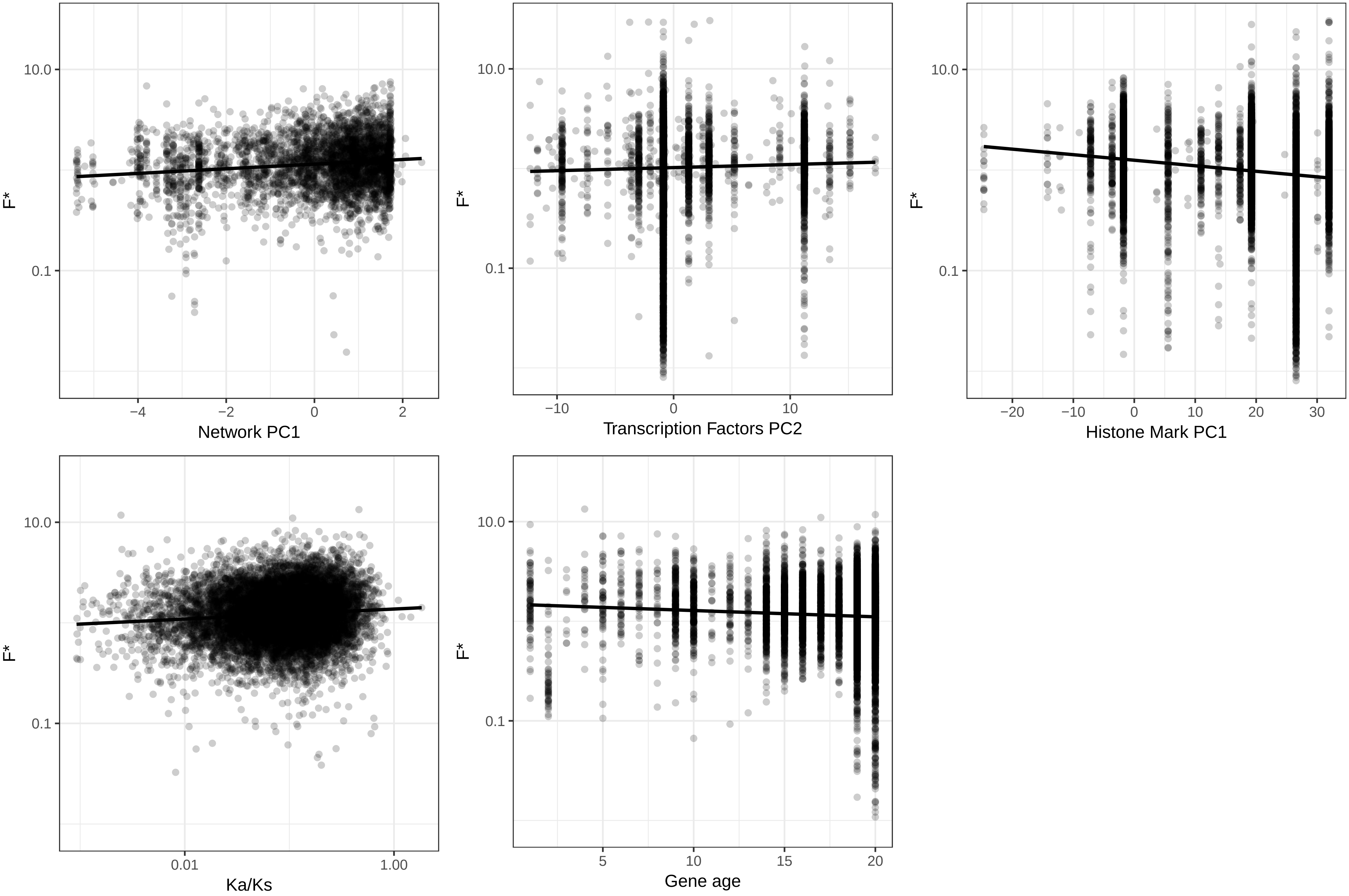
Figure S4: Logistic principal component analysis of histone marks. A) Proportion of deviance explained by models with 1, 2, etc. principal components. B) Loadings of each variable on the 2 first components.

Figure S5: Assortativity in networks. Assortativity for F* and hub score are plotted against each other. Orange line: simple linear model. Blue line: “breakpoint” model. Vertical dashed line show the minimal value of hub score assortativity from which it has no effect on F* assortativity.


Interactome
Protein
DEGRADATION
TRANSLATION
Expression noise
R
AAAAAA
Messenger
DEGRADATION
Transcription noise
TRANSCRIPTION
Regulome
TFs
E
P
CDS
E
Gene


A
Variance
Fano factor
Noise
F*
SGE measure
1e+01
1e+03
1e+05
1e+07
1
100
10.0
0.1
1
100
10000
Mean
B
density
0.0
0.5
1.0
0.1
10.0
F*


Cell Cycle
Translation
Eukaryotic Translation Initiation
Cap−dependent Translation Initiation
GTP hydrolysis and joining of the 60S ribosomal subunit
3' −UTR−mediated translational regulation
L13a−mediated translational silencing of Ceruloplasmin expression
Signaling by Rho GTPases
SRP−dependent cotranslational protein targeting to membrane
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
Nonsense−Mediated Decay (NMD)
Formation of a pool of free 40S subunits
Cellular responses to stress
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
Chromatin organization
Chromatin modifying enzymes
Eukaryotic Translation Elongation
Peptide chain elongation
Eukaryotic Translation Termination
The citric acid (TCA) cycle and respiratory electron transport
Processing of Capped Intron−Containing Pre−mRNA
mRNA Splicing
mRNA Splicing − Major Pathway
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Downstream signaling events of B Cell Receptor (BCR)
HATs acetylate histones
Regulatory RNA pathways
Respiratory electron transport
Activation of the mRNA upon binding of the cap−binding complex and eIFs, and subsequent binding to 43S
Ribosomal scanning and start codon recognition
Translation initiation complex formation
Formation of the ternary complex, and subsequently, the 43S complex
Transcriptional regulation by small RNAs
Epigenetic regulation of gene expression
Cellular response to heat stress
Regulation of HSF1−mediated heat shock response
mRNA Splicing − Minor Pathway
0.02
0.04
0.06
0.08
0.10
GeneRatio
Count
20
30
40
50
p.adjust
0.008
0.006
0.004
0.002


F*
10.0
0.1
−4
−2
0
2
Network PC1
F*
10.0
0.1
−10
0
10
Transcription Factors PC2
F*
10.0
0.1
−20
−10
0
10
20
30
Histone Mark PC1
F*
10.0
0.1
0.01
1.00
Ka/Ks
F*
10.0
0.1
5
10
15
20
Gene age