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

noise

AUTHORS: Gustavo Valadares Barroso¹; Natasa Puzovic¹ 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, typically 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 take 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 show 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.

34 Introduction

35 Isogenic cell populations display phenotypic variability even in homogeneous environments
 36 (Spudich and Koshland 1976). This observation challenged the clockwork view of the intra-cellular
 37 molecular machinery and led to the recognition of the stochastic nature of gene expression. Since
 38 biochemical reactions result from the interactions of individual molecules in small numbers
 39 (Gillespie 1977), the inherent stochasticity of binding and diffusion processes generates noise along
 40 the biochemical cascade leading to the synthesis of a protein from its encoding gene (**Figure 1**).
 41 The study of stochastic gene expression (SGE) classically recognizes two sources of expression
 42 noise. Following the definition introduced by Elowitz et al (Elowitz et al. 2002), extrinsic noise
 43 results from variation in concentration, state and location of shared key molecules involved in the
 44 reaction cascade from transcription initiation to protein folding. This is because molecules that are
 45 shared among genes, such as ribosomes and RNA polymerases, are typically present in low copy
 46 numbers relative to the number of genes actively transcribed (Shahrezaei and Swain 2008).
 47 Extrinsic factors also include physical properties of the cell such as size and growth rate, likely to
 48 impact the diffusion process of all molecular players. Extrinsic factors therefore affect every gene in
 49 a cell equally. Conversely, intrinsic factors generate noise in a gene-specific manner. They involve,
 50 for example, the strength of cis-regulatory elements (Suter et al. 2011) as well as the stability of the
 51 mRNA molecules that are transcribed (Mcadams and Arkin 1997; Thattai and Oudenaarden 2001).
 52 Every gene is affected by both sources of stochasticity and the relative importance of each has been
 53 discussed in the literature (Becskei et al. 2005; Raj and Oudenaarden 2008). Shahrezaei and Swain
 54 (Shahrezaei and Swain 2008) proposed a more general, systemic and explicit definition for any
 55 organization level, where intrinsic stochasticity is “generated by the dynamics of the system from
 56 the random timing of individual reactions” and extrinsic stochasticity is “generated by the system
 57 interacting with other stochastic systems in the cell or its environment”. This generic definition
 58 therefore includes Raser and O’Shea’s (Raser and O’Shea 2005) suggestion to further distinguish
 59 extrinsic noise occurring “within pathways” and “between pathways”. Other organization levels of
 60 gene expression are also likely to affect expression noise, such as chromatin structure (Blake et al.
 61 2003; Hebenstreit 2013), and three-dimensional genome organization (Pombo and Dillon 2015).

62 Pioneering work by Fraser et al (Fraser et al. 2004) has shown that SGE is an evolvable trait
 63 which is subject to natural selection. First, genes involved in core functions of the cell are expected
 64 to behave more deterministically (Barkai and Leibler 1999) because temporal oscillations in the
 65 concentration of their encoded proteins are likely to have a deleterious effect. Second, genes
 66 involved in immune response (Arkin et al. 1998; Norman et al. 2015) and response to
 67 environmental conditions can benefit from being unpredictably expressed in the context of selection

68 for bet-hedging (Thattai and Oudenaarden 2004). As the relation between fitness and stochasticity
69 depends on the function of the underlying gene, selection on SGE is expected to act mostly at the
70 intrinsic level (Newman et al. 2006; Lehner 2008; Wang and Zhang 2011). The molecular
71 mechanisms by which natural selection operates to regulate expression noise, however, remain to be
72 elucidated.

73 Due to methodological limitations, seminal studies on SGE (both at the mRNA and protein
74 levels) have focused on only a handful of genes (Elowitz et al. 2002; Ozbudak et al. 2002; Chubb et
75 al. 2006). The canonical approach consists in selecting genes of interest and recording the change of
76 their noise levels in a population of clonal cells as a function of either (1) the concentration of the
77 molecule that allosterically controls affinity of the transcription factor to the promoter region of the
78 gene (Blake et al. 2003; Bar-even et al. 2006) or (2) mutations artificially imposed in regulatory
79 sequences (Ozbudak et al. 2002). In parallel with theoretical work (Kepler and Elston 2001; Batada
80 and Hurst 2007; Kaufmann and van Oudenaarden 2007; Sánchez and Kondev 2008), these
81 pioneering studies have provided the basis of our current understanding of the proximate molecular
82 mechanisms behind SGE, namely complex regulation by transcription factors, architecture of the
83 upstream region (including the presence of TATA box) and gene orientation (Wang et al. 2011),
84 translation efficiency and mRNA / protein stability (Eldar and Elowitz 2010), properties of the
85 protein-protein interaction network (Li et al. 2010). Measurements at the genome scale coupled
86 with rigorous statistical analyses are however needed in order to go beyond gene idiosyncrasies
87 and particular histories, and test hypotheses about the evolutionary forces shaping SGE (Sauer et al.
88 2007).

89 The recent advent of single-cell RNA sequencing makes it possible to sequence the
90 transcriptome of each individual cell in a collection of clones, and to observe the variation of gene-
91 specific mRNA quantities across cells. This provides a genome-wide assessment of transcriptional
92 noise. While not accounting for putative noise resulting from the process of translation of mRNAs
93 into proteins, transcriptional noise accounts for noise generated by both synthesis and degradation
94 of mRNA molecules (**Figure 1**). Previous studies, however, have shown that transcription is a
95 limiting step in gene expression, and that transcriptional noise is therefore a good proxy for
96 expression noise (Newman et al. 2006; Taniguchi et al. 2011). Here, we used publicly available
97 single-cell transcriptomics data sets to quantify gene-specific transcriptional noise and relate it to
98 other genomic factors, including protein conservation and position in the interaction network, in
99 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 FPKM² 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 performed polynomial regressions in the log-space plot of variance *versus* mean. We defined F^* as $\sigma_{obs}^2/\sigma_{pred}^2$ (see Material and Methods) that is, the ratio of the observed variance over the variance component predicted by the mean expression level. We selected the simplest model for which no correlation between F^* and mean expression was observed, and found that a degree 3 polynomial model was sufficient to remove further correlation (Kendall's tau = -0.0037, p-value = 0.5217, **Figure 2A**). 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 2B**). This approach fulfills the same goal as the running median approach of Newman et. al (Newman et al. 2006), whilst it includes the effect of mean expression directly into the measure of stochasticity instead of correcting a posteriori a dependent measure (in that case, the Fano factor). We therefore use F^* as a measure of SGE throughout this study.

131 **Stochastic gene expression correlates with the three-dimensional** 132 **structure of the genome**

133 We first sought to investigate whether genome organization significantly impacts the
134 patterns of stochastic gene expression. We assessed whether genes in proximity along chromosomes
135 display more similar amount of transcriptional noise than distant genes. We tested this hypothesis
136 by computing the primary distance on the genome between each pair of genes, that is, the number
137 of base pairs separating them on the chromosome, as well as the relative difference in their
138 transcriptional noise (see Methods). We found no significant association between the two distances
139 (Mantel tests, each chromosome tested independently). Contiguous genes in one dimension,
140 however, have significantly more similar transcriptional noise than non-contiguous genes
141 (permutation test, p -value $< 1e-4$, **Figure S1**). Using Hi-C data from mouse embryonic cells (Dixon
142 et al. 2012), we report that genes in contact in three-dimensions have significantly more similar
143 transcriptional noise than genes not in contact (permutation test, p -value $< 1e-3$, **Figure S1**). Most
144 contiguous genes in one-dimension also appear to be close in three-dimensions and the effect of 3D
145 contact is stronger than that of 1D contact. These results therefore suggest that the three-
146 dimensional structure of the genome has a stronger impact on stochastic gene expression than the
147 position of the genes along the chromosomes. We further note that while highly significant, the size
148 of this effect is small, with a difference in relative expression of -1.10% (**Figure S1**).

149 **Transcription factors binding and histone methylation impact** 150 **stochastic gene expression**

151 The binding of transcription factors (TF) to promoter constitutes one notable source of
152 transcriptional noise (**Figure 1**) (Blake et al. 2003; Newman et al. 2006). In eukaryotes, the
153 accessibility of promoters is determined by the chromatin state, which is itself controlled by histone
154 methylation. We assessed the extent to which transcriptional noise is linked to particular TFs and
155 histone marks by using data from the Ensembl regulatory build (Zerbino et al. 2015), which
156 provides data from experimental evidence of TF binding and methylation sites along the genome.
157 First we contrasted the F^* values of genes with binding evidence for each annotated TF
158 independently. Among 13 TF represented by at least 5 genes in our data set, we found that 4 of
159 them significantly influence F^* after adjusting for a global false discovery rate of 5%: the
160 transcription repressor CTCF (adjusted p -value = 0.0321), the transcription factor CP2-like 1
161 (Tcfcp2l1, adjusted p -value = 0.0087), the X-Linked Zinc Finger Protein (Zfx, adjusted p -value =
162 0.0284) and the Myc transcription factor (MYC, adjusted p -value = 0.0104). Interestingly,
163 association with each of these four TFs led to an increase in transcriptional noise. We also report a

164 weak but significant positive correlation between the number of transcription factors associated
165 with each gene and the amount of transcriptional noise (Kendall's tau = 0.0238, p-value = 0.0007).
166 This observation is consistent with the idea that noise generated by each TF is cumulative (Sharon
167 et al. 2014). We then tested if particular histone marks are associated with transcriptional noise.
168 Among five histone marks represented in our data set, three were found to be highly significantly
169 associated to a higher transcriptional noise: H3K4me3 (adjusted p-value = 1.9981e-146), H3K4me2
170 (adjusted p-value = 5.4524e-121) and H3K27me3 (adjusted p-value = 5.2985e-34). Methylation on
171 the fourth Lysine of histone H3 is associated with gene activation in humans, while tri-methylation
172 on lysine 27 is usually associated with gene repression (Barski et al. 2007). These results suggest
173 that both gene activation and silencing contribute to the stochasticity of gene expression, in
174 agreement with the view that bursty transcription leads to increased noise (Blake et al. 2003;
175 Newman et al. 2006; Batada and Hurst 2007).

176 **Low noise genes are enriched for housekeeping functions**

177 We investigated the function of genes at both ends of the F^* spectrum. We defined as
178 candidate gene sets the top 10% least noisy or the top 10% most noisy genes in our data set, and
179 tested for enrichment of GO terms and Reactome pathways (see Methods). It is expected that genes
180 encoding proteins participating in housekeeping pathways are less noisy because fluctuations in
181 concentration of their products might have stronger deleterious effects (Pedraza and van
182 Oudenaarden 2005). On the other hand, stochastic gene expression could be selectively
183 advantageous for genes involved in immune and stress response, as part of a bet-hedging strategy
184 (eg Arkin et al. 1998; Shalek et al. 2013). GO terms enrichment test revealed significant categories
185 enriched in the low noise gene set only: molecular functions "nucleic acid binding" and "structural
186 constituent of ribosome", the biological processes "nucleosome assembly", "innate immune
187 response in mucosa" and "translation", as well as the cellular component "cytosolic large ribosomal
188 subunit" (**Table 1**). All these terms but one relate to gene expression, in agreement with previously
189 reported findings in yeast (Newman et al. 2006). We further find a total of 41 Reactome pathways
190 significantly over-represented in the low-noise gene set (false discovery rate set to 1%).
191 Interestingly, the top most significant pathways belong to modules related to translation (RNA
192 processing, initiation of translation and ribosomal assembly), as well as several modules relating to
193 gene expression, including chromatin regulation and mRNA splicing (**Figure 3**). Only one pathway
194 was found to be enriched in the high noise set: TP53 regulation of transcription of cell cycle genes
195 (p-value = 0.0079). This finding is interesting because TP53 is a central regulator of stress response
196 in the cell (Hussain and Harris 2006). These results therefore corroborate previous findings that
197 genes involved in stress response might be evolving under selection for high noise as part of a bet

208 hedging strategy (Shalek et al. 2013; Viney and Reece 2013). The small amount of significantly
209 enriched Reactome pathways by high noise genes can potentially be explained by the nature of the
210 data set: as the original experiment was based on unstimulated cells, genes that directly benefit from
211 high SGE might not be expressed in these experimental conditions.

202 **Highly connected proteins are synthesized by low-noise genes**

203 The structure of the interaction network of proteins inside the cell can greatly impact the
204 evolutionary dynamics of genes (Jeong et al. 2000; Barabási and Oltvai 2004). Furthermore, the
205 contribution of each constitutive node within a given network varies. This asymmetry is largely
206 reflected in the power-law-like degree distribution that is observed in virtually all biological
207 networks (Barabási and Albert 1999) with a few genes displaying many connections and a majority
208 of genes displaying only a few. The individual characteristics of each node in a network can be
209 characterized by various measures of centrality (Newman 2003). Following previous studies on
210 protein evolutionary rate (Fraser et al. 2002; Hahn et al. 2004; Jovelín and Phillips 2009) and
211 protein-protein interaction (PPI) networks (Li et al. 2010) we asked whether, at the gene level, there
212 is a link between centrality of a protein and the amount of transcriptional noise. We study six
213 centrality metrics measured on two types of network data: (i) pathway annotations from the
214 Reactome database (Fabregat et al. 2016) and (ii) PPI data from the iRefIndex database. PPI data
215 are typically more complete (5,553 genes with gene expression data) but do not provide functional
216 evidence. The Reactome database is based on published functional evidence, but encompasses less
217 genes (4,454 genes for which expression data is available). In addition, graph representing PPI
218 network are not oriented while graph representing Pathway annotations are, implying that distinct
219 statistics can be computed on both types of networks.

220 We first estimated the pleiotropy index of each gene by counting how many different
221 pathways the corresponding proteins are involved in. We then computed centrality measures as
222 averages over all pathways in which each gene is involved. These measures include (1) *node*
223 *degree*, which corresponds to the number of other nodes a given node is directly connected with, (2)
224 *hub score*, which estimates the extent to which a node links to other central nodes, (3) *authority*
225 *score*, which estimates the importance of a node by assessing how many hubs link to it, (4)
226 *transitivity*, or *clustering coefficient*, defined as the proportion of neighbors that also connect to
227 each other, (5) *closeness*, a measure of the topological distance between a node and every other
228 reachable node (the fewer edge hops it takes for a protein to reach every other protein in a network,
229 the higher its closeness), and (6) *betweenness*, a measure of the frequency with which a protein
230 belongs to the shortest path between every pairs of nodes.

231

232 We find that node degree, hub score, authority score and transitivity are all significantly
 233 negatively correlated with transcriptional noise on pathway-based networks: the more central a
 234 protein is, the less transcriptional noise it displays (**Figure 4A-D** and **Table 2**). We also observed
 235 that pleiotropy is negatively correlated with F^* (Kendall's tau = -0.0514, p-value = 8.31E-07,
 236 **Figure 4E**, **Table 2**), suggesting that a protein that potentially performs multiple functions at the
 237 same time needs to be less noisy. This effect is not an artifact of the fact that pleiotropic genes are
 238 themselves more central (e.g. correlation of pleiotropy and node degree: Kendall's tau = 0.2215, p-
 239 value < 2.2E-16) or evolve more slowly (correlation of pleiotropy and Ka / Ks ratio: Kendall's tau =
 240 -0.1060, p-value < 2.2E-16) since it is still significant after controlling for these variables (partial
 241 correlation of pleiotropy and F^* , accounting for centrality measures and Ka / Ks: Kendall's tau =
 242 -0.0254, p-value = 7.45E-06). Closeness and betweenness, on the other hand, show a negative
 243 correlation with F^* , yet much less significant (Kendall's tau = -0.0254, p-value = 0.0109 for
 244 closeness and tau = -0.0175, p-value = 0.0865 for betweenness, see **Figure 4FG** and **Table 2**). In
 245 modular networks (Hartwell et al. 1999) nodes that connect different modules are extremely
 246 important to the cell (Guimera and Amaral 2005) and show high betweenness scores. In yeast, high
 247 betweenness proteins tend to be older and more essential (Joy et al. 2005), an observation also
 248 supported by our data set (betweenness vs gene age, Kendall's tau = 0.0619, p-value = 1.09E-07;
 249 betweenness vs Ka/Ks, Kendall's tau = -0.0857, p-value = 3.83E-16). It has been argued, however,
 250 that in protein-protein interaction networks high betweenness proteins are less essential due to the
 251 lack of directed information flow, compared to, for instance, regulatory networks (Yu et al. 2007), a
 252 hypothesis which could explain the observed lack of correlation.

253 By applying similar measures on the PPI network, we report significant negative correlation
 254 between F^* and PPI centrality measures (**Figure 4H-K**, **Table 2**). Because the PPI network is not
 255 directed, authority scores and hub scores cannot be distinguished. The results obtained with the
 256 mouse PPI interaction network are qualitatively similar to the ones obtained by Li et al (2010) on
 257 Yeast expression data (Li et al. 2010). In addition, we further report that genes involved in complex
 258 interactions (that is, genes which interact with more than one other protein simultaneously) have
 259 reduced noise in gene expression (Wilcoxon rank test, p-value = 8.053E-05, **Figure 4L**),
 260 corroborating previous findings in Yeast (Fraser et al. 2004). Conversely, genes involved in
 261 polymeric interactions, that is, where multiple copies of the encoded protein interact with each
 262 other, did not show significantly different noise than other genes (Wilcoxon rank test, p-value =
 263 0.0821, **Figure 4M**).

264 It was previously shown that centrality measures negatively correlate with evolutionary rate
 265 (Hahn and Kern 2004). Our results suggest that central genes are selectively constrained for their
 266 transcriptional noise, and that centrality therefore also influences the regulation of gene expression.

267 Interestingly, it has been reported that central genes tend to be more duplicated (Vitkup et al. 2006).
 268 The authors proposed that such duplication events would have been favored as they would confer
 269 greater robustness to deleterious mutations in proteins. Our results are compatible with another, non
 270 exclusive, possible advantage: having more gene copies could reduce transcriptional noise by
 271 averaging the amount of transcripts produced by each gene copy (Raser and O'Shea 2005).

272 **Network structure impacts transcriptional noise of constitutive genes**

273 Whereas estimators of node centrality highlight gene-specific properties inside a given
 274 network, measures at the whole-network level enable the comparison of networks with distinct
 275 properties. We computed the size, diameter and global transitivity for each annotated network in our
 276 data set (1,364 networks, Supplementary Material) which we compare with the average F^* measure
 277 of all constitutive nodes. The size of a network is defined as its total number of nodes, while
 278 diameter is the length of the shortest path between the two most distant nodes. Transitivity is a
 279 measure of connectivity, defined as the average of all nodes' clustering coefficients. Interestingly,
 280 while network size is positively correlated with average degree and transitivity (Kendall's tau =
 281 0.5880, p-value < 2.2e-16 and Kendall's tau = 0.1166, p-value = 1.08E-10, respectively), diameter
 282 displays a positive correlation with average degree (Kendall's tau = 0.2959, p-value < 2.2e-16) but
 283 a negative correlation with transitivity (Kendall's tau = -0.0840, p-value = 2.17E-05). This is
 284 because diameter increases logarithmically with size, that is, addition of new nodes to large
 285 networks do not increase the diameter as much as additions to small networks. This suggests that
 286 larger networks are relatively more compact than smaller ones, and their constitutive nodes are
 287 therefore more connected. We find that average transcriptional noise correlates negatively with
 288 network size (Kendall's tau = -0.0514, p-value = 0.0039), while being independent of the diameter
 289 (Kendall's tau = 0.0061, p-value = 0.7547 see **Table 3**). These results are in line with the node-
 290 based analyses, and show that the more connections a network has, the less stochastic the
 291 expression of the underlying genes is. This supports the view of Raser and Oshea (Raser and
 292 O'Shea 2005) that the gene-extrinsic, pathway-intrinsic level is functionally pertinent and needs to
 293 be distinguished from the globally extrinsic level. We further asked whether genes with similar
 294 transcriptional noise tend to synthesize proteins that connect to each other (positive assortativity) in
 295 a given network, or on the contrary, tend to avoid each other (negative assortativity). We considered
 296 all Reactome pathways annotated to the mouse and estimated their respective F^* assortativity. We
 297 found the mean assortativity to be significantly negative, with a value of -0.1384 (one sample
 298 Wilcoxon rank test, p-value < 2.2e-16), meaning that proteins with different F^* values tend to
 299 connect with each other (**Figure S3**). Maslov & Sneppen (Maslov and Sneppen 2002) reported a
 300 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 significantly positive (average of 0.1221, one sample Wilcoxon rank test, p -value = 1.212×10^{-12} , **Figure S5**), 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 = 0.2581, p -value < 2.2×10^{-16}). The relationship between the measures, however, is not linear (**Figure S5**), 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 = 0.2730, p -value < 2.2×10^{-16}), while for pathways with positive hub score assortativity the effect vanishes (Kendall's tau = 0.0940, p -value = 3.135×10^{-4}). 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 = -0.0818, one sample Wilcoxon test with p -value < 2.2×10^{-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 (K_a) to synonymous substitutions (K_s), 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 K_a / K_s ratio and F^* (**Figure 4N**, Kendall's tau = 0.0557, p -value < 1.143×10^{-5}), that is, highly constrained genes (low K_a / K_s ratio) display less transcriptional noise (low F^*) than fast evolving ones. This result demonstrates that genes encoding proteins under strong purifying selection are also more constrained on their transcriptional noise.

334 Older genes are less noisy

335 Evolution of new genes was long thought to occur via duplication and modification of
 336 existing genetic material (“evolutionary tinkering”, (Jacob 1977)). Evidence for *de novo* gene
 337 emergence is however becoming more and more common (Tautz and Domazet-Lošo 2011; Xie et
 338 al. 2012). *De novo* created genes undergo several optimization steps, including their integration into
 339 a regulatory network (Neme and Tautz 2013). We tested whether the historical process of
 340 incorporation of new genes into pathways impacts the evolution of transcriptional noise. We used
 341 the phylostratigraphic approach of Neme & Tautz (Neme and Tautz 2013), which categorizes genes
 342 into 20 strata, to compute gene age and tested for a correlation with F*. As older genes tend to be
 343 more conserved (Wolf et al. 2009), more central (according to the preferential attachment model of
 344 network growth (Jeong et al. 2000; Jeong et al. 2001)) and more pleiotropic, we controlled for these
 345 confounding factors (Kendall's tau = -0.0663, p-value = 1.58E-37 ; partial correlation controlling
 346 for Ka / Ks ratio, centrality measures and pleiotropy level, **Figure 4O**). These results suggest that
 347 older genes are more deterministically expressed while younger genes are more noisy. While we
 348 cannot rule out that functional constraints not fully accounted for by the Ka / Ks ratio or unavailable
 349 functional annotations could explain at least partially the correlation of gene age and transcriptional
 350 noise, we hypothesise that the observed correlation results from ancient genes having acquired more
 351 complex regulation schemes through time. Such schemes include for instance negative feedback
 352 loops, which have been shown to stabilize gene expression and reduce expression noise (Beckstein
 353 and Serrano 2000; Thattai and Oudenaarden 2001).

354 Position in the protein network is the main driver of transcriptional 355 noise

356 In order to jointly assess the effect of network topology, epigenomic factors, Ka / Ks ratio
 357 and gene age, we modeled the patterns of transcriptional noise as a function of multiple predictive
 358 factors within the linear model framework. This analysis could be performed on a set of 2,794 genes
 359 for which values were available jointly for all variables. In order to avoid colinearity issues because
 360 some of these variables are intrinsically correlated, we performed data reduction procedures prior to
 361 modeling. For continuous variables, including Pathway and PPI network variables, Ka / Ks ratio
 362 and gene age, we conducted a principal component analysis (PCA) and used as synthetic measures
 363 the first eight principal components (PC), explaining together more than 80% of the total inertia
 364 (**Figure S2A**). The first principal component (PC1) of the PCA analysis is associated with pathway
 365 centrality measures (degree, hub score, authority score and transitivity, **Figure S2B**). The second
 366 principal component (PC2) corresponds to PPI centrality measures (degree, hub score and

betweenness), while the third component (PC3) relates to gene age and Ka / Ks ratio. The fourth component (PC4) is associated with PPI complex interactions and transitivity. PC5 and PC6 are essentially associated to betweenness and closeness of the pathway network, PC7 with PPI polymeric interactions and PC8 with pathway pleiotropy. 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 (hereby noted TFPC), which explained 78% of deviance (**Figure S3A**). The loads on the first component (TFPC1) are all negative, meaning that TFPC1 captures a global correlation trend and does not discriminate between TFs. Tcfcp2l1 appears to be the TF with the highest correlation to TFPC1. The second component TFPC2 is dominated by TCFC (positive loading) and Oct4 (negative loading), while the third component TFPC3 is dominated by Esrrb (positive loading) and MYC, nMyc and E2F1 (negative loadings, **Figure S3B**). For histone marks, the two first components (hereby noted HistPC) explained 95% of variance and were therefore retained (**Figure S4A**). HistPC1 is dominated by marks H3K27me3 linked to gene repression (negative loadings) and HistPC2 by marks H3K4me1 and H3K4me3 linked to gene activation (positive loadings, **Figure S4A**).

We fitted a linear model with F* as a response variable and all 13 synthetic variables as explanatory variables. We find that PC1 has a significant positive effect on F* (**Table 3**). As the loadings of the centrality measures on PC1 are negative (**Figure S2C**), this result is consistent with our finding of a negative correlation of pathway-based centrality measure with F*. PC3 has a highly significant negative effect on F*, which is consistent with a negative correlation with gene age (positive loading on PC3) and a positive correlation with the Ka / Ks ratio (negative loading on PC3, **Figure S2D**). The last highly significant variable is the first principal component of the logistic PCA on histone methylation patterns, HistPC1, which has a negative effect on F*. Because the loadings are essentially negative on HistPC1, this suggests a positive effect of methylation, in particular the repressive H3K27me3. Altogether, the linear model with all variables explained 4.01% of the total variance (adjusted R²). 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. To compare the individual effects of each explanatory variable, we conducted a relative importance analysis. As a mean of comparison, we fitted a similar model with mean expression as a response variable. We find that pathway centrality measures (PC1 variables) account for 38% of the explained variance, while protein constraints and gene age (PC3) account for 32%. Chromatin state (HistPC1) account for another 15% of the variance (**Figure 5**). These results contrast with the model of mean expression, where HistPC1 and HistPC2 respectively account for 51% and 9% of the explained variance, and PC1 and PC3 20%

402 and 10% only (**Figure 5**). This suggests (1) that among all factors tested, position in protein
403 network is the main driver of the evolution of gene-specific stochastic expression, followed by
404 protein constraints and gene age and (2) that different selective pressures act on the mean and cell-
405 to-cell variability of gene expression.

406 We further included the effect of three-dimensional organization of the genome in order to
407 assess whether it could act as a confounding factor. We developed a correlation model allowing for
408 genes in contact to have correlated values of transcriptional noise. The correlation model was fitted
409 together with the previous linear model in the generalized least square (GLS) framework. This
410 model allows for one additional parameter, λ , which captures the strength of correlation due to
411 three-dimensional organization of the genome (see Methods). The estimate of λ was found to be
412 0.0016, which means that the spatial autocorrelation of transcriptional noise is low on average. This
413 estimate is significantly higher than zero, and model comparison using Akaike's information
414 criterion favors the linear model with three-dimensional correlation (AIC = 4880.858 vs. AIC =
415 4890.396 for a linear model without three-dimensional correlation). Despite the significant effect of
416 3D genome correlation, our results were qualitatively and quantitatively very similar to the model
417 ignoring 3D correlation (**Table 3**).

418 **Analysis of bone marrow-derived dendritic cells supports the** 419 **generality of the results**

420 We assessed the reproducibility of our results by analyzing an additional single-cell
421 transcriptomics data set of 95 unstimulated bone marrow-derived dendritic cells (BMDC) (Shalek et
422 al. 2014). After filtering (see Methods), the data set consisted of 11,640 genes. Using the same
423 normalization procedure as for the ESC data set, we nonetheless report a weak but significant
424 negative correlation between F^* and the mean expression, even with a degree-5 polynomial
425 regression (-0.0459 , $p\text{-value} < 1.13\text{E-}13$). This effect is due to the distribution of per-gene, between
426 cell RFKM values being extremely skewed in this data set. In order to assess the impact of the
427 residual correlation with the mean, we computed a value of F^* (noted F_R^*) on a restricted dataset
428 where the variance was between 1/8 and 8 times the mean (75% of all genes) using a quantile
429 regression on the median instead of a linear regression. A second degree polynomial quantile
430 regression proved to be sufficient to remove the effect of mean expression (Kendall's tau = 0.0114,
431 $p\text{-value} = 0.1125$) on this restricted data set. As all results were consistent when using the F_R^* and
432 F^* measures, we only discuss here results obtained with F^* and refer to **Supplementary Data 1** for
433 detailed results obtained with the F_R^* measure.

434 We report a highly significant positive correlation between F^* values measured on the 8,792
 435 genes with expression in both data sets, suggesting that cell-to-cell variance in gene expression is to
 436 a large extent conserved among the two cell types (Kendall's tau = 0.1289, p-value < 2.2E-16,
 437 **Figure S6A**). GO terms or reactome pathways enrichment analyses reveal less significant but
 438 consistent terms with the ESM analysis: the high F^* gene set did not show any significantly
 439 enriched GO term or reactome pathway (FDR set to 1%) and the low F^* gene set revealed RNA-
 440 binding as a significantly enriched molecular function, as well as 21 enriched pathways (**Figure**
 441 **S7**). In agreement with results from the ESM analysis, many of the most significant enriched
 442 pathways relate to gene expression, including translation and splicing. Interestingly, the two most
 443 significant pathways, however, are "Vesicle-mediated transport" and "Membrane trafficking", two
 444 essential pathways for the functioning of dendritic cells. Analyses of network centrality measures
 445 also generally show consistent results with the ESC data set, more central genes displaying reduced
 446 gene expression noise (**Figure S6B-N, Table S1**). Quantitative differences consists of PPI
 447 betweenness, as well as pathway closeness and betweenness are highly significantly negatively
 448 correlated with F^* while they were only weakly or non-significant with the ESC data set. The only
 449 discrepancies that we report between the two data sets relate to pathway level statistics. Pathway
 450 size appears to be significantly positively correlated with mean F^* , while it was negatively
 451 correlated on the ESC data set, yet with a comparatively higher p-value. Similarly pathway diameter
 452 is significantly positively correlated with mean F^* in the BMDC data set, while it was not
 453 significant with the ESC data. We currently have no hypothesis to explain this particular
 454 discrepancy. While these results support the generality of our observations, they also illustrate that
 455 in details, the fine structure of translational noise may vary in a cell type-specific manner.

456 We fitted linear models as for the embryonic stem cell (ESC) data set, with the exception
 457 that no epigenomic data was available for this cell type. Data reduction was performed using a
 458 principal component analysis, with the eight first principal components explaining 81% of the total
 459 deviance (**Figure S8A**). We report consistent results with the ESC analysis, with all major effects
 460 similar in direction and intensity, highlighting the impact of network centrality measures on
 461 expression noise (**Table S2**). With the BMDC data, however, the second principal component PC2
 462 which is associated with PPI centrality measures (**Figure S8B**) appears to have a significant
 463 negative impact on F^* , while it was not significant with the ESC dataset. As the loading of the PPI
 464 centrality measures are positive on PC2, this is consistent with central genes having a lower
 465 transcriptional noise as for the pathway network metrics (**Figure S8C**). When taking 3D genome
 466 correlations into account, we estimated a low correlation coefficient as for the ESC dataset (λ
 467 = 0.0004), and the AIC favored the model without correlation in this case. Relative importance
 468 analysis revealed that network centrality measures contributed most to the explained variance (48%

and 21% for PC1 and PC2 respectively), while the contribution of protein constraints and gene age (PC3) was 24%.

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_t^* value for each gene (see Methods). We fitted linear models as for the single cell data, using F_t^* instead of F^* . We report that no variable had a significant effect on F_t^* (Table S3). In addition, there was no enrichment of the lower 10th F_t^* percentile for any particular pathway or GO term. The upper 90th percentile showed no GO term enrichment, but four pathways appeared to be significant: “Chromosome maintenance” (adjusted p-value = 0.0043), “Polymerase switching on the C-strand of the telomere” (adjusted p-value = 0.0062), “Polymerase switching” (adjusted p-value = 0.0062) and “Leading strand synthesis” (adjusted p-value = 0.0062), which relate to DNA replication. While it is unclear why genes involved in these pathways would display higher technical variance in RNA sequencing, these results strikingly differ from our analyses of single cell RNA sequencing and therefore suggest that technical variance does not act as a confounding factor in our analyses.

Because only three replicates were available in the pooled RNA-Seq data set, we asked whether the resulting estimate of mean and variance in expression is accurate enough to allow proper inference of noise and its correlation with other variables. We conducted a jackknife procedure where we sampled the original cells from the ESC data set and re-estimated F^* for each sample. We tested combinations of 3, 5, 10 and 15 cells, with 1,000 samples in each case. In each sample, we computed F^* with the same procedure as for the complete data set, and fitted a linear model with all 13 synthetic variables. For computational efficiency, we did not include 3D

correlation in this analysis. We compute for each variable the number of samples where the effect is significant at the 5% level and has the same sign as in the model fitted on the full data set. We find that the model coefficients are very robust to the number of cells used (**Figure S9A**) and that 3 cells are enough to infer the effect of the PC1 and PC3 variables, the most significant in our analyses. Two main conclusions can be drawn from this jackknife analysis: (1) that the lack of significant effect of our explanatory variables on technical noise is not due to the low number of replicates used to compute the mean and variance in expression and (2) that our conclusions are very robust to the actual cells used in the analysis, ruling out drop-out and amplification biases as possible source of errors (Kharchenko et al. 2014).

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 error rates. (4) While the newly introduced F^* measure allowed us to assess the distribution of transcriptional noise independently of the average mean expression, 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.

535 In a pioneering study, Fraser et al (Fraser et al. 2004), followed by Shalek et al (Shalek et al.
536 2013), demonstrated that essential genes whose deletion is deleterious, and genes encoding subunits
537 of molecular complexes as well as housekeeping genes display reduced gene expression noise. Our
538 findings go beyond these early observations by providing a statistical assessment of the joint effect
539 of multiple explanatory factors. Our analyses reveal that network centrality measures are the
540 explanatory factors that explained the most significant part of the distribution of transcriptional
541 noise in the genome. Network-based statistics were first tested by Li et al (Li et al. 2010) using PPI
542 data in Yeast. While we are able to extend these results to mouse cells, we show that more detailed
543 annotation as provided by the Reactome database lead to new insights into the selective forces
544 acting on expression noise. Our results suggest that “pathways” constitute a relevant systemic level
545 of organisation, at which selection can act and drive the evolution of SGE at the gene level. This
546 multi-level selection mechanism, we propose, can be explained by selection against noise
547 propagation within networks. It has been experimentally demonstrated that expression noise can be
548 transmitted from one gene to another gene with which it is interacting (Pedraza and van
549 Oudenaarden 2005). Large noise at the network level is deleterious (Barkai and Leibler 1999) but
550 each gene does not contribute equally to it, thus the strength of selective pressure against noise
551 varies among genes in a given network. We have shown that highly connected, “central” proteins
552 typically display reduced transcriptional noise. Such nodes are likely to constitute key players in the
553 flow of noise in intra-cellular networks as they are more likely to transmit noise to other
554 components. In accordance with this hypothesis, we find genes with the lowest amount of
555 transcriptional noise to be enriched for top-level functions, in particular involved in the regulation
556 of other genes.

557 These results have several implications for the evolution of gene networks. First, this means
558 that new connections in a network can potentially be deleterious if they link genes with highly
559 stochastic expression. Second, distinct selective pressures at the “regulome” and “interactome”
560 levels (**Figure 1**) might act in opposite direction. We expect genes encoding highly connected
561 proteins to have more complex regulation schemes, in particular if their proteins are involved in
562 several biological pathways. In accordance, several studies demonstrated that expression noise of a
563 gene positively correlates with the number of transcription factors controlling its regulation (Sharon
564 et al. 2014), a correlation that we also find significant in the data set analyzed in this work. Central
565 genes, while being under negative selection against stochastic behavior, are then more likely to be
566 controlled by numerous transcription factors which increase transcriptional noise. As a
567 consequence, if the number of connections at the interactome level is correlated with the number of
568 connections at the regulome level, we predict the existence of a trade-off in the number of
569 connections a gene can make in a network. Alternatively, highly connected genes might evolve

570 regulatory mechanisms allowing them to uncouple these two levels: negative feedback loops, for
571 instance, where the product of a gene down-regulates its own production have been shown to
572 stabilize expression and significantly reduce stochasticity (Becskei and Serrano 2000; Dublanche et
573 al. 2006; Tao et al. 2007). We therefore predict that negative feedback loops are more likely to
574 occur at genes that are more central in protein networks, as they will confer greater resilience
575 against high SGE, which is advantageous for this class of genes.

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

590 Conclusion

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

600 **Material and Methods**

601 **Single-cell gene expression data set**

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

610 **Measure of transcriptional noise**

611 The expression mean (μ) and variance (σ^2) of each gene over all single cells were
 612 computed. We measured stochastic gene expression as the ratio $F^* = \frac{\sigma^2}{\sigma^2(\mu)}$, where $\sigma^2(\mu)$ is
 613 the expected variance given the mean expression. In order to compute $\sigma^2(\mu)$, we performed
 614 several polynomial regressions with $\log(\sigma^2)$ as a function of $\log(\mu)$, with degrees between 1
 615 and 5. We then tested the resulting F^* measures for residual correlation with mean expression using
 616 Kendall's rank correlation test. We find that a degree-3 polynomial regression was sufficient to
 617 remove any residual correlation with F^* (Kendall's tau = 0.0037, p-value = 0.5217). F^* can be seen
 618 as a general expression for the Fano factor and noise measure: when using a polynome of degree 1,
 619 the expression of F^* becomes $F^* = \frac{\sigma^2}{\exp(a+b \cdot \log(\mu))} = \frac{\sigma^2}{\exp(a) \cdot \mu^b}$, and is therefore equivalent to
 620 the Fano factor when $a = 0$ and $b = 1$, and equivalent to noise when $a = 0$ and $b = 2$.

621 **Genome architecture**

622 The mouse proteome from Ensembl (genome version: mm9) was used in order to get
 623 coordinates of all genes. The Hi-C dataset for embryonic stem cells (ES) from Dixon et al (Dixon et
 624 al. 2012) was used to get three-dimensional domain information. Two genes were considered in
 625 proximity in one dimension (1D) if they are on the same chromosome and no protein-coding gene
 626 was found between them. The primary distance (in number of nucleotides) between their midpoint
 627 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 10,000 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. 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 (binary model). Alternatively, $\delta_{i,j}$ can be defined as the actual number of contacts between the two 20 kb regions (as defined by Dixon et al) the genes belong to (proportional model). Parameter λ was estimated jointly with other model parameters, it measures the strength of the genome “spatial” correlation. Models were compared using Akaike’s information criterion (AIC). We find that the proportional correlation model fitted the data better and therefore selected it for further analyses.

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.

659 **Biological pathways, protein-protein interactions and network** 660 **topology**

661 We defined genes either in the top 10% least noisy or in the top 10% most noisy as
662 candidate sets and used the Reactome PA package (Yu and He 2016) to search the mouse Reactome
663 database for overrepresented pathways with a 1% false discovery rate.

664 Centrality measures were computed using a combination of the “igraph” (Csardi and Nepusz
665 2006) and “graphite” (Sales et al. 2012) packages for R. As the calculation of assortativity does not
666 handle missing data (that is, nodes of the pathway for which no value could be computed), we
667 computed assortativity on the sub-network with nodes for which data were available. Reactome
668 centrality measures could be computed for a total of 4,454 genes with expression data.

669 Protein-protein interactions (PPI) were retrieved from the iRefIndex database (Razick et al.
670 2008) using the iRefR package for R (Mora and Donaldson 2011). Interactions were converted to a
671 graph using the dedicated R functions in the package, and the same methods were used to compute
672 centrality measures as for the pathway analysis. Because the PPI-based graph was not oriented,
673 authority scores were not computed for this data (as this gave identical results to hub scores).
674 Furthermore, as most genes are part of a single graph structure in the case of PPI interactions,
675 closeness values were not further analysed as they were virtually identical for all genes.

676 **Gene Ontology Enrichment**

677 Eight thousand three hundreds and twenty five out of the 13,660 genes were associated with
678 Gene Ontology (GO) terms. We tested genes for GO terms enrichment at both ends of the F*
679 spectrum using the same threshold percentile of 10% low / high noise genes as we did for the
680 Reactome analysis. We carried out GO enrichment analyses using two different algorithms:
681 “Parent-child” (Grossmann et al. 2007) and “Weight01”, a mixture of two algorithms developed by
682 Alexa et al (Alexa et al. 2006). We kept only the terms that appeared simultaneously on both
683 Parent-child and Weight01 under 1% significance level, controlling for multiple testing using the
684 FDR method (Benjamini and Hochberg 1995).

685 **Sequence divergence**

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

689 Gene Age

690 The relative taxonomic ages of the mouse genes have been computed and is available in the
691 form of 20 Phylostrata (Neme and Tautz 2013). Each Phylostratum corresponds to a node in the
692 phylogenetic tree of life. Phylostratum 1 corresponds to “All cellular organisms” whereas
693 Phylostratum 20 corresponds to “*Mus musculus*”, with other levels in between. We used this
694 published information to assign each of our genes to a specific Phylostratum and used this as a
695 relative measure of gene age: Age = 21 - Phylostratum, so that an age of 1 corresponds to genes
696 specific to *M. musculus* and genes with an age of 20 are found in all cellular organisms.

697 Linear modeling

698 We simultaneously assessed the effect of different factors on transcriptional noise by fitting
699 linear models to the gene-specific F^* estimates. To avoid colinearity issues of intrinsically
700 correlated explanatory variables, we conducted a data reduction procedure using multivariate
701 analysis. We used variants of principal component analysis (PCA) on explanatory variables in three
702 groups: network centrality measures, K_a / K_s and gene age with standard PCA, transcription factor
703 binding evidence and histone methylation patterns using logistic PCA, a generalization of PCA for
704 binary variables (Landgraf and Lee 2015). In each case, we used the most representative
705 components (totaling at least 75% of the total deviance) as synthetic variables. PCA analysis was
706 conducted using the ade4 package for R (Dray and Dufour 2007), logistic PCA was performed
707 using the logisticPCA package (Landgraf and Lee 2015).

708 We built a linear model with F^* as a response variable and thirteen synthetic variables as
709 explanatory variables. As the synthetic variables are principal components, they are orthogonal by
710 construction. The fitted model displayed significant departure to normality and was further
711 transformed using the Box-Cox procedure (“boxcox” function from the MASS package for R
712 (Venables and Ripley 2002)). Residues of the selected model had normal, independent residue
713 distributions (Shapiro-Wilk test of normality, p-value = 0.121, Ljung-Box test of independence, p-
714 value = 0.2061) but still displayed significant heteroscedasticity (Harrison-McCabe test, p-value =
715 0.003). In order to ensure that this departure from the Gauss-Markov assumptions does not bias our
716 inference, we used the “robcov” function of the “rms” package in order to get robust estimates of
717 the effect significance (Harrell 2015). Relative importance of each explanatory factor was
718 assessed using the method of Lindeman, Merenda and Gold (Lindeman et al. 1979) as implemented
719 is the R package “relaimpo”. The significance of the level of variance explained by each factor was
720 computed using standard ANOVA procedure.

721 **Additional data sets**

722 The aforementioned analyses were additionally conducted on the bone marrow-derived
723 dendritic cells data set of Shalek et al (Shalek et al. 2014). Following the filtering procedure
724 established by the authors in the original paper, genes which did not satisfied the condition of being
725 expressed by an amount such that $\log(\text{TPM}+1) > 1$ in at least one of the 95 single cells were further
726 discarded, where TPM stands for transcripts per million. This cut-off threshold resulted in 11,640
727 genes being kept for investigation. The rest of the analyses was conducted in the same way as for
728 the ESM data set.

729 **Jackknife procedure**

730 A jackknife procedure was conducted in order to assess (1) the robustness of our results to
731 the choice of actual cells used to estimate mean and variance in gene expression and (2) the power
732 of the pooled RNA-seq analysis for which only three replicates were available. This analysis was
733 conducted by sampling 3, 5, 10 and 15 of the original 20 single cells of the ESM data set (Sasagawa
734 et al. 2013), 1,000 times in each case. The exact same analysis was conducted on each random
735 sample as for the complete data set, and model coefficients and their associated p-values were
736 recorded.

737 **Data and program availability**

738 All datasets and scripts to reproduce the results of this study are available under the DOI
739 10.6084/m9.figshare.4587169.

741 **Authors contributions**

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

744 **Acknowledgements**

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

750

751 **References**

- Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22:1600–1607.
- 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.
- Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* 39:945–949.
- 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.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.
- 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.
- Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* 1695:1695.
- 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.
- Dray S, Dufour A-B. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* [Internet] 22. Available from: <http://www.jstatsoft.org/v22/i04/>

- 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.
- Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, K€orninger F, McKay S, et al. 2016. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* 44:D481–487.
- 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.
- Gillespie DT. 1977. Exact Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry* 81:2340–2361.
- Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* 23:3024–3031.
- 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.
- Harrell FE. 2015. Regression Modeling Strategies. Cham: Springer International Publishing Available from: <http://link.springer.com/10.1007/978-3-319-19425-7>
- 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.
- Hussain SP, Harris CC. 2006. p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *J Nippon Med Sch* 73:54–64.
- 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.
- Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11:740–742.
- 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.
- Li J, Min R, Vizeacoumar FJ, Jin K, Xin X, Zhang Z. 2010. Exploiting the determinants of stochastic gene expression in *Saccharomyces cerevisiae* for genome-wide prediction of expression noise. *Proc. Natl. Acad. Sci. U.S.A.* 107:10472–10477.
- Lindeman RH, Merenda PF, Gold RZ. 1979. *Introduction to Bivariate and Multivariate Analysis*. Glenview, Ill: Scott Foresman & Co
- 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.
- Mora A, Donaldson IM. 2011. iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics* 12:455.
- 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.
- Newman MEJ. 2003. The Structure and Function of Complex Networks. *SIAM Review* 45:167–256.
- 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 applied, poorly understood. *BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS* . 6:141–148.
- Raj A, Oudenaarden AV. 2008. Review Nature , Nurture , or Chance : Stochastic Gene Expression and Its Consequences. *Cell* 135:216–226.
- Raser JM, O’Shea EK. 2005. Noise in Gene Expression: Origins, Consequences, and Control. *Science* 309.
- Razick S, Magklaras G, Donaldson IM. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9:405.
- Sales G, Calura E, Cavalieri D, Romualdi C. 2012. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* 13:20.
- 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.

- 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.
- Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. New York, NY: Springer New York Available from: <http://link.springer.com/10.1007/978-0-387-21706-2>
- Viney M, Reece SE. 2013. Adaptive noise. *Proc Biol Sci* [Internet] 280. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3735249/>
- 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 G-Z, Lercher MJ, Hurst LD. 2011. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol Evol* 3:320–331.
- 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 G, He Q-Y. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* 12:477–479.
- 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:0003735	structural constituent of ribosome	2.28E-07	6.81E-20
MF	GO:0003676	nucleic acid binding	8.16E-06	6.06E-04
BP	GO:0006412	translation	4.08E-08	7.15E-12
BP	GO:0002227	innate immune response in mucosa	6.49E-04	6.22E-03
CC	GO:0022625	cytosolic large ribosomal subunit	4.48E-03	1.40E-12

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, as estimated from pathway annotations and protein-protein interactions networks.

Data	Measure	Correlation with F*	p-value
Pathways	Degree	-0.0745	1.14E-13 ***
	Hub score	-0.0808	6.61E-16 ***
	Authority score	-0.0666	2.72E-11 ***
	Clustering coefficient	-0.0794	4.55E-15 ***
	Closeness	-0.0254	1.09E-02 *
	Betweenness	-0.0175	8.65E-02 .
	Pleiotropy	-0.0514	8.31E-07 ***
	Size	-0.0514	3.91E-03 ***
	Diameter	0.0061	7.55E-01 NS
	Global transitivity	-0.1532	3.06E-17 ***
PPI	Degree	-0.0249	8.20E-03 **
	Hub score	-0.0942	< 2.2E-16 ***
	Transitivity	-0.0338	6.24E-04 ***
	Betweenness	-0.0140	1.31E-01 NS

Note: All correlations are computed using Kendall's rank correlation test, with p-value codes defined as *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1. NS = non-significant. PPI: protein-protein interactions.

764

765 **Table 3:** Linear models of transcriptional noise with genomic and epigenomic factors.

	OLS			GLS		
	Coefficient	SE	p-value	Coefficient	SE	p-value
(Intercept)	0.1612	0.0781	0.0392 *	0.1665	0.0663	0.0121 *
PC1	0.0390	0.0065	<0.0001 ***	0.0396	0.0065	0.0000 ***
PC2	-0.0048	0.0069	0.4854	-0.0048	0.0069	0.4838
PC3	-0.0526	0.0091	<0.0001 ***	-0.0518	0.0092	0.0000 ***
PC4	-0.0102	0.0097	0.2905	-0.0109	0.0100	0.2773
PC5	0.0117	0.0106	0.2713	0.0123	0.0106	0.2456
PC6	-0.0152	0.0107	0.1536	-0.0152	0.0109	0.1623
PC7	0.0210	0.0102	0.0384 *	0.0211	0.0110	0.0561 .
PC8	0.0100	0.0113	0.3778	0.0073	0.0114	0.5250
TFPC1	0.0028	0.0041	0.4912	0.0025	0.0034	0.4658
TFPC2	0.0025	0.0027	0.3664	0.0024	0.0026	0.3585
TFPC3	0.0032	0.0042	0.4513	0.0032	0.0037	0.3825
HistPC1	-0.0031	0.001	0.0015 **	-0.0033	0.0010	0.0007 ***
HistPC2	-0.0027	0.0016	0.0846 .	-0.0029	0.0015	0.0566 .

766

767 Note: OLS: Ordinary Least Squares. GLS: Generalized Least Squares. SE: standard error. Pathway
768 PC1-8: principal components on centrality measures, protein conservation and gene age. TFPC1-3:
769 principal components of the logistic PCA on transcription factors binding evidences. HistPC1 and
770 2: principal components of the logistic PCA on histone modification marks.

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: Variance, Fano factor (variance / mean), noise (square of the coefficient of variation, variance / mean²) and F* (this study). Lines represent quantile regression fits (median, first and third quartiles). Point and bars represent median, first and third quartiles for each category of mean expression obtained by discretization of the x axis. B) Distribution of F* over all genes in this study. Vertical line corresponds to F* = 1.

Figure 3: Enriched pathways in the low-noise gene set. Depicted pathways are the fifteen most significant in the 10% genes with lowest transcriptional noise.

Figure 4: Factors driving stochastic gene expression. Correlation of F* and all tested network centrality measures, as well as protein conservation (Ka / Ks ratio) and gene age. Point and bars represent median, first and third quartiles for each category of mean expression obtained by discretization of the x axis, together with the quantile regression lines estimated on the full data set.

Figure 5: Relative importance of explanatory factors on mean gene expression and F*. Significance codes refer to ANOVA test of variance, *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1.

Supplementary material:

Table S1: Correlation of transcriptional noise with genes centrality measures and pleiotropy for the bone marrow-derived dendritic cells data set. Legends as in **Table 2**.

Table S2: Linear models of transcriptional noise with genomic factors for the bone marrow-derived dendritic cells data set. Legend as in Table 4.

Table S3: Linear model 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 10,000 permutations (see Methods). Left panel: distribution for neighbor genes along the genome. Right panel: distribution for genes in contact in three-dimensions.

800 **Figure S2:** Principal component analysis of pathways centrality measures. A) Proportion of
801 deviance explained by models with 1, 2, etc. principal components. B) Contributions, computed as
802 proportion of deviance, of each input variable to each principal component. C) Loadings of each
803 variable on the 2 first components. D) Loadings of each variable on the 3rd and 4th principal
804 components.

805 **Figure S3:** Logistic principal component analysis of transcription factor binding evidences. A)
806 Proportion of deviance explained by models with 1, 2, etc. principal components. B) Contributions,
807 computed as proportion of deviance, of each input variable to each principal component. C)
808 Loadings of each variable on the 2 first components. D) Loadings of each variable on the 2nd and
809 3rd principal components.

810 **Figure S4:** Logistic principal component analysis of histone marks. A) Proportion of deviance
811 explained by models with 1, 2, etc. principal components. B) Contributions, computed as proportion
812 of deviance, of each input variable to each principal component. C) Loadings of each variable on
813 the 2 first components.

814 **Figure S5:** Assortativity in networks. A) Distribution of assortativity values for hub scores. B)
815 Distribution of assortativity values for F*. C) Assortativity for F* and hub scores are plotted against
816 each other. Solid lines represent linear regressions fitted on pathways with negative or positive hub
817 score assortativity, respectively. Dashed line represents a linear regression fitted on all data.

818 **Figure S6:** Factors driving stochastic gene expression in the bone marrow-derived dendritic cells
819 data set. Legends as in **Figure 4**.

820 **Figure S7:** Enriched pathways in the low-noise gene set of the bone marrow-derived dendritic cells
821 data set.

822 **Figure S8:** Principal component analysis of pathways centrality measures of the bone marrow-
823 derived dendritic cells data set. Legends as in **Figure S2**.

824 **Figure S9:** Robustness and power analysis. A jackknife procedure was conducted by fitted linear
825 models with all explanatory variables on a subset of cells taken randomly (x-axis). A) estimated
826 coefficient of each effect. B) proportion of simulations where the coefficient is significant at the 5%
827 level. Filled bars correspond to significant effect when the complete data set is used. PC: principal
828 component. PPI: protein-protein interactions. TF: transcription factors.

829

830 **Supplementary Data 1:** All scripts and data set necessary to reproduce the analyses and figures in
831 this manuscript.

Expression noise

Transcription noise

Factor

Variable

F*

Mean