Gene expression variability - the other dimension in transcriptome analysis

de Jong, Tristan V; Moshkin, Yuri M; Guryev, Victor

Published in:
Physiological Genomics

DOI:
10.1152/physiolgenomics.00128.2018

Publication date:
2019

Article type: Review

Title: **Gene expression variability – the other dimension in transcriptome analysis**

Authors: Tristan V. de Jong[1], Yuri M. Moshkin[2,3], Victor Guryev[1]

Affiliations:

[1] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands.

[2] Institute of Cytology and Genetics, Siberian Branch of RAS, Novosibirsk, Russia

[3] Institute of Molecular and Cellular Biology, Siberian Branch of RAS, Novosibirsk, Russia

Corresponding authors:

Dr. Yuri M Moshkin; Tel: +41 76 699 42 04; E-mail: yury.moshkin@gmail.com

Prof. Victor Guryev; Tel: +31 6 5272 4873; E-mail: v.guryev@umcg.nl

Abstract (215 words)

Transcriptome sequencing is a powerful technique to study molecular changes that underlie the differences in physiological conditions and disease progression. A typical question that is posed in such studies is finding genes with significant changes between sample groups. In this respect expression variability is regarded as a nuisance factor that is primarily of technical origin and complicates the data analysis. However, it is becoming apparent that the biological variation in gene expression might be an important molecular phenotype that can affect physiological parameters.

In this review we explore the recent literature on technical and biological variability in gene expression, sources of expression variability, (epi-) genetic hallmarks and evolutionary constraints in genes with robust and variable gene expression. We provide an overview of recent findings on effects of external cues, such as diet and aging on expression variability and on other biological phenomena that can be linked to it. We discuss metrics and tools that were developed for quantification of expression variability and highlight the importance of future studies in this direction.

In order to assist the adoption of expression variability analysis, we also provide a detailed description and computer code, which can easily be utilized by other researchers. We also provide a re-analysis of recently published data to highlight the value of the analysis method.

Main text

Affordable sequencing has greatly advanced our understanding of changes in transcription programs and their relation to diseases. One of the sequencing-enabled technologies, transcriptome profiling by RNA-sequencing (RNA-seq) is becoming increasingly popular to study molecular phenotypes. The main advantages of this method, when compared to hybridization microarray-based approaches, include an increased sensitivity and larger dynamic range, its ability to detect unannotated transcripts and transcript isoforms and, importantly, it enables digital quantification (counting) of RNA molecules. As a result, RNA-seq has the potential to quantify lowly expressed genes, to reveal subtle changes in gene expression (115), to discover new genes, transcript isoforms and allelic variants for proteogenomics analysis (53), and, as will be discussed later, digital quantification of RNAs simplifies statistical analysis of gene expression and interpretation of its variability.

The typical analysis of RNA-sequencing data focuses on the finding of genes that show differential expression between groups. Such analysis can be done with tools like edgeR (58) or DEseq2 (52). The results call attention to genes which significantly change with respect to an average RNA copy number between measurable factors like age, diet, the knock-down/-out/-in of genes of interest, and so on. Unfortunately, in such analysis, variability in gene expression is often ignored as it is treated as a nuisance that only diminishes statistical power. At the same time, gene expression is naturally a stochastic process and in some cases its fluctuation, rather than the mean RNA copy number, could be significantly influenced by an experimental factor or a physiological state. Thus, while variations caused by technical factors can be considered as the true nuisance factor (80), differential variability in gene expression caused by biological factors might represent a layer of information on gene regulation just as important as changes in the mean expression levels (104). In this review we discuss recent studies exploring gene expression fluctuations, their approach to quantification of expression variability, contribution to understanding of the principles underlying physiological homeostasis and potential to uncover additional molecular phenotypes associated with disease.

Sources of variability in gene expression: Poisson - "intrinsic" vs non-Poisson - "extrinsic" gene noise.

The inter-sample differences among transcriptome profiles originate from biological events as well as from experimental procedures. The latter represents a source of technical noise due to the collection and storage of samples, the isolation of RNA, selection of RNA molecules and the preparation of library (92). Library amplification and sequencing might also introduce differences depending on instruments, read length and mode of sequencing. All these factors have potential to complicate the analysis of biological variability in gene expression, especially for large (inter-) national and prospective projects where data is being produced using different versions of instruments and/or kits (58). Thus, when studying variation in gene expression, it is important to estimate technical variability through comparison of technical replicates prepared from the same starting material (111) and compare it to the degree of variability seen among biologically different samples (58).

Putting technical variability aside, gene noise originates from the stochastic nature of chemical reactions driving RNA synthesis (birth) and degradation (death). In a stationary state and in the absence of upstream cellular drives, a process of RNA "birth-death" is expected to be a stochastic Poisson process (21, 96). This process is described by just two kinetic parameters, namely the synthesis (λ) and degradation (γ) rates. The expectation (mean) and variance of RNA copy number are given by the Poisson rate ($E[\mathrm{RNA}] = Var[RNA] = \mu$) represented by a constant ratio of synthesis to degradation rates: $\mu = \frac{\lambda}{\gamma} = \hat{\lambda}$. Gene expression noise, expressed through a squared coefficient of variation in RNA copy number, is reciprocal to the mean of RNA copy number: $cv^2(\mathrm{RNA}) = \frac{Var[RNA]}{E[\mathrm{RNA}]^2} = \mu^{-1}$ (96). Here, we will refer to this as Poisson noise following (21, 66, 96). However, in reality gene synthesis is more complex as it is regulated by so-called upstream cellular drives (21). Because upstream cellular drives are stochastic themselves, the RNA "birth-death" becomes a doubly stochastic (mixed) Poisson process. Consequently, this increases the gene expression noise to the amount that is contributed by all upstream drives, which we will refer to as non-Poisson noise following (21, 66, 96).

For example, promoter switching between active (ON) and inactive (OFF) states acts as such a drive (Fig. 1). The probability of the promoter to be in ON state ($\mathrm{p_{on}}$) is a Beta-distributed random variable, which depends on RNA degradation rate normalized $\hat{\mathrm{k}}_{\mathrm{on}} = \frac{\mathrm{k_{on}}}{\gamma}$ and $\hat{\mathrm{k}}_{\mathrm{off}} = \frac{\mathrm{k_{off}}}{\gamma}$ rates of promoter switching: $\mathrm{p_{on}} \sim Beta(\hat{\mathrm{k}}_{\mathrm{on}}, \hat{\mathrm{k}}_{\mathrm{off}})$. This, in turn, defines the

distribution of otherwise constant Poisson rate ($\mu = \hat{\lambda}\mathrm{p_{on}}$) as Beta-Poisson (21, 72). A convenient property of mixed Poisson distributed random variables is that they allow for simple derivation of their moments (expectation and variance) just from the moments of the mixing distribution (44). That is $E[\mathrm{RNA}] = \hat{\lambda}E[\mathrm{p_{on}}] = \langle\mu\rangle$ and $Var[\mathrm{RNA}] = \langle\mu\rangle + Var[\mu] = \langle\mu\rangle + \langle\mu\rangle^2 Var[\mathrm{p_{on}}]$, from where $cv^2(\mathrm{RNA}) = \langle\mu\rangle^{-1} + cv^2(\mu) = \langle\mu\rangle^{-1} + cv^2(\mathrm{p_{on}})$ (Fig. 1). Thus, the total gene noise sums from Poisson noise ($\langle\mu\rangle^{-1}$) and non-Poisson noise caused by upstream cellular drive, namely promoter switching ($cv^2(\mu) = cv^2(\mathrm{p_{on}})$).

Under limiting conditions of $\hat{k}_{\mathrm{off}} \gg \hat{k}_{\mathrm{on}}$ and $\hat{k}_{\mathrm{off}} \gg 1$, i.e. when a gene is transcribed in short bursts, the $\mathrm{p_{on}}$ distribution converges to Gamma ($\mathrm{p_{on}} \sim Gamma(\hat{\mathrm{k}}_{\mathrm{on}}, \hat{\mathrm{k}}_{\mathrm{off}})$) and the resulting distribution of RNA copy number is Gamma-Poisson (also known as Negative-Binomial)(72). The Gamma-Poisson representation helps understanding of how Poisson and non-Poisson noise are related to often measured in single cell experiments parameters of transcription, namely the burst size (a number of molecules synthesized in a burst) and burst frequency (93). That is because Poisson noise equals to $\langle\mu\rangle^{-1} = (bf_b)^{-1}$ and non-Poisson noise is $cv^2(\mu) = cv^2(\mathrm{p_{on}}) = f_b^{-1}$, where $b = \lambda k_{\mathrm{off}}^{-1}$ is a burst size and $f_b = \hat{k}_{\mathrm{on}}$ is a burst frequency (21, 72). Thus, non-Poisson noise is inversely related to burst frequency, which implies that changes in burst frequency are indicative of changes in non-Poisson noise. For the detailed derivations of various stochastic gene expression models under a mixed Poisson framework and further theoretical insights we refer to a compelling work by Dattani and Barahona (21).

In essence, the partitioning of the total gene noise into Poisson and non-Poisson, immediately corresponds to a concept of “intrinsic” and “extrinsic” gene noise (26, 94). Two-colour reporter gene assays allow for the separation of within cell variations from cell-to-cell variation in gene expression. In this assay two identical copies of a promoter drive the expression of reporters: yellow fluorescent protein (YFP) and green fluorescent protein (GFP). The single-cell fluorescence readout will show different expression levels of YFP and GFP due to the intrinsically stochastic nature of gene expression. At the same time extrinsic noise will be related to covariance between expression levels of these two reporters. Hence, the within cell gene expression fluctuations have been coined as “intrinsic” gene noise, while cell-to-cell variations were referred to as “extrinsic” gene noise. A total gene noise

sums, then, from “intrinsic” and “extrinsic” resulting in identical partitioning of noise as Poisson and non-Poisson.

However, defining gene noise through a combination of “intrinsic” and “extrinsic” noise has been subjected to criticism. First, it is not clear relative to what within biological system gene noise is “intrinsic” or “extrinsic” (68). Second, they are conditioned on each other (88). Indeed, “intrinsic” gene noise, or Poisson noise for that matter, is reciprocal to the mean gene expression. For the two-state promoter model, i.e. in the presence of upstream cellular drive caused by promoter fluctuation, the mean gene expression depends on the probability of the promoter to be in the ON state. Thus, “intrinsic” gene noise is coupled to upstream cellular drives. Likewise, “extrinsic” gene noise depends on the RNA lifetime normalized rates of promoter switching between the ON and OFF states. Thus, “extrinsic” gene noise is conditioned on the characteristic gene state variables (21, 72).

Having this in mind and considering that RNA “birth-death” is a doubly stochastic Poisson process, it makes more sense to stay with Poisson and non-Poisson partitioning of gene expression noise (21). Accordingly, parameters affecting the gene expression variability and thus the gene expression noise, could be classified into gene state variables (kinetic parameters of RNA synthesis/degradation), regulatory variables (concentration of transcription factors, chromatin accessibility, epigenetic state, etc.) and system state variables (aging, metabolism or other environmental factors acting on cells) (Fig. 1).

Gene state determinants of expression variability.

If the right conditions are met, RNA polymerase Pol II (RNAP II) binds to a promoter region and initiates transcription of the gene (81). The transcription happens in short bursts followed by a refractory period in which no transcription takes place (116). A simplified derivation of the two-state promoter model shows that non-Poisson noise depends inversely on the burst frequency, while Poisson noise is reciprocal of a product of burst size and burst frequency (21, 72). Each gene has its own bursting dynamics which, in turn, determines its noise (93). Different factors can either influence the burst frequency, a frequency of promoter activation within the mean lifetime of RNA, or the burst size, the amount of RNA produced per unit of time within a burst (82). Thus, any factor interfering

155 with promoter fluctuation, RNA synthesis or degradation kinetics is expected to modulate
156 the within-cell and cell-to-cell variability in RNA copy number and thus gene noise.

157 In eukaryotes, the RNA "birth-death" rates are orchestrated by a complex regulatory
158 system. With respect to the regulation of the synthesis rate, it is worth mentioning the RNA
159 splicing process. The different proteins involved in splicing and accessibility of alternative
160 donor/acceptor sites can modulate RNAP II elongation rate and, thus, the RNA synthesis
161 rate. For instance, RNAP II elongation rates tend to increase throughout introns as
162 compared to exons (42, 46). However, splice sites themselves, in mammals, but not in yeast,
163 act as RNAP II pausing sites (19, 41). Such pausing can be bypassed by the inhibition of
164 splicing mechanisms (65). To that, co-transcriptional checkpoints associated with splicing
165 can further modulate the synthesis rates (3, 16). Thus RNA splicing, being intimately linked
166 with RNA elongation, is expected to contribute to Poisson noise by modulating RNA "birth"
167 rate.

168 The amount of RNA observed in a cell is the consequence of equilibrium between
169 synthesis and degradation. This means not only fluctuations in the synthesis rate, but also
170 variations in the degradation rate are likely to influence both the average expression as well
171 as the variation in expression (57, 97). The half-life of RNA molecules depends on the length
172 of the 3'-poly(A)-tail which is removed through deadenylation prior to degradation (67,
173 109). As a direct consequence of the two-state promoter model, the total gene expression
174 noise (Poisson and non-Poisson) is directly proportional to the RNA degradation rate. This
175 implies an increased noise level for RNA species with shorter half-life and a decreased noise
176 for the stable RNA molecules. For example, the presence of certain microRNAs have been
177 shown to increase the rate of RNA deadenylation (107) and one can predict that such a
178 mechanism will turn up the gene noise. Strikingly, although RNA synthesis and
179 degradation, at first glance, are two independent processes, the RNA degradation rate was
180 found to be regulated by transcription (13, 33). In terms of gene noise, the existence of a
181 coupling between synthesis and degradation rates has a profound consequence as it leads
182 to non-Poisson RNA "birth-death" process even in the absence of upstream cellular drives
183 (96).

Finally, it is reasonable to assume that the kinetics of transcriptional bursts and as a result, gene noise is likely to be determined by the promoter sequence and the surrounding architecture. Indeed, the presence of a TATA-box within the promoter is known to increase not only the average expression of genes, but also its noise (11, 76, 77). TATA-box binding protein TBP associates with distinct co-activator complexes, each of which competes for the binding to the promoter, as it also mediates re-initiation of transcription by RNAP II (77, 81). Consequently, this promotes fluctuations in promoter activity, i.e. increases cell-to-cell or temporal deviations in the probability of the promoter to be in ON state. This, in turn, increases the gene expression noise, as non-Poisson noise is directly related to the fluctuations in these upstream cellular drives (21). Likewise, the complexity of the promoter can further increase the competition between distinct transcription factors and the expression noise. A simple promoter architecture, in which the promoter region is free from secondary regulation tends to generate little noise (36, 87). DNA variants in the promoter region can modulate the binding affinity of transcription factors, consequently changing both the average gene expression and expression noise (36). Besides competition for transcription factor binding within a promoter, competition between distinct promoters might also increase the gene noise by lowering the promoter burst frequencies (77). Next to that, the presence of a so called 'speed bumps' downstream of the transcription start site can cause RNAP II stalling (1), which might be detrimental for determination of bursting kinetics and noise. Although, further mechanistic insights into impact of gene state variables on gene noise remain to be made, the logic of doubly stochastic Poisson "birth-death" process and the two-state promoter model provide valuable tools for the dissection of gene noise determinants through the modelling of RNA "birth-death" rates.

Epigenetic determinants of expression variability.

In eukaryotes, promoter accessibility and RNA synthesis are modulated by the epigenetic state of a gene, which sums from the DNA methylation status, nucleosome assembly and post-translational histone modifications. The epigenetic landscape is not static, as environmental cues such as diet, smoking, physical exercises and ageing can alter the epigenetic composition of the chromatin throughout organism's lifetime (8, 29, 34, 95, 102). Methylation patterns were shown to vary with circadian rhythm (5). Methylation of

CpG islands in promoter regions can alter transcription dynamics, resulting in the repression of transcription (10). In general, the presence of CpG islands in promoters lowers gene noise (27, 60). This might seem somewhat paradoxical, as increased CpG methylation is associated with increased nucleosome occupancy (20) and, as result, it is expected to elevate gene noise because of the lower promoter accessibility for transcription factor binding. However, occurrence of CpG islands in promoters of robustly expressed genes, i.e. in genes with low transcriptional noise, does not imply an increased methylation of their promoters. At the same time, a long-standing hypothesis suggests that DNA methylation might suppress cryptic transcription initiation from within the body of a gene, thereby reducing transcriptional noise (9, 39). Thus, it will be important to address these factors in future research on how DNA methylation partitions between promoter and gene body in genes with robust and noisy expression.

Assembly of eukaryotic DNA into nucleosomes adds yet another layer of complexity to gene regulation and is likely to modulate gene expression noise (17). Indeed, TATA-containing promoters favouring nucleosome assembly tend to further increase the noise due to a competition between TBP and nucleosomes (18, 83). Further, histones that constitute nucleosomes are subjected to a wide range of post-translational modifications, collectively known as a histone code (4). Transcription co-activator or co-repressor complexes recognize particular combinations of histone modifications tuning both gene expression level and expression variability (27, 108, 112). Thus, it may not be surprising that the presence of conflicting histone marks, i.e. co-occurrence of histone modifications associated with gene activation and repression, leads to an increased expression variability (27). First, bivalent histone modifications are expected to create a competitive state at the promoter and as a result, increase noise in the promoter activation. Second, bivalent histone marks might interfere with transcription initiation and elongation causing RNAP II to pause (51). In general, increased acetylation of histones and an overall “loose” chromatin structure at the promoter are associated with low expression noise, whilst a “closed” chromatin configuration and deficiency in active histone marks drive a higher noise (14, 22, 63, 90, 98). In conclusion, the stochastic nature of RNA synthesis is intimately modulated by the stochastic nature of chromatin and DNA methylation states acting as upstream cellular drives (14, 28).

System state determinants of expression variability.

In general, the biological processes are affected by two time-dependent factors: the circadian clock and aging. Interestingly, gene expression variability is also linked to these factors. For example, recently it has been shown that the circadian clock modulates burst frequency rather than burst size. Consequently, gene expression noise oscillates daily along with the average gene expression (63). Aging deteriorates many physiological parameters and their variability increases with time (reviewed in 15) and a clear epigenetic drift between monozygotic twins arises during aging (29). Thus, aging is expected to have a profound effect on gene expression variability (55). In accordance with this, the expression of housekeeping genes was shown to be more robust in cardiomyocytes from young mice as compared to old mice (6). To that, recent studies in mouse models provide evidence that inter-individual variability in gene expression tends to increase with age and can be reduced upon interventions aimed to slow ageing (61, 105). Further, a lower variation in gene expression was observed to correlate with the presence of H3K36me3 (27), a histone mark that was previously associated with increased longevity (86), although it is not known whether this epigenetic modification is a cause or consequence of the increased variation in gene expression. A recent study of gene expression in human skin, fat and blood samples from twin samples showed a general decrease of expression variability with age of individuals studied (101), This surprising and, perhaps, contradictory observation on linking aging and expression variability warrant further investigations of expression variability using other populations, tissue types as well as computational approaches for its quantification.

Variability in gene expression might explain many biological phenomena.

Variability determines plasticity, i.e. a degree to which a gene can change its expression in response to environmental fluctuations as a consequence of fluctuation-response relationship (49, 84). Plasticity of expression can serve a cell to adapts to a new environment (106). At the population level, a more varied expression of certain genes can produce individuals that are better prepared for changing conditions at the cost of reduced metabolic efficiency (12). This was shown on a microscopic scale in yeast, in which a high variability in expression of yeast plasma-membrane transporters enhanced their adaptive capabilities to a changing environment (114). Selection for the yeast TDH3 enzyme involved in the glucose metabolism was shown to have a greater impact on expression noise rather

than on the average level of expression, showing an example of selection for higher variability as an adaptation mechanism (59). Overall, genes involved in environmental responses show more variation in expression, which can be beneficial for non-housekeeping functions such as coping with stress or reacting to environmental queues (11, 69). Genome wide analysis of transcriptional and epigenetic variability across human immune cell types showed that neutrophils, one of the first-responder cells upon an infection, contained the largest variation in both methylation and expression and alluding that variability might be an important factor in immune response (24). Also inter-population variability has shown that genes can have similar levels of expression variability across individuals and populations, with the largest differences observed among genes associated with immune response and disease susceptibility such as chemokine receptor *CRCX4* that is important for HIV-1 infection, where variation in expression may underlie difference in disease susceptibility (50). In contrast, genes involved in growth and development (85), as well as genes which provide a universal function, such as protein synthesis or degradation generally (e.g. translation initiation and ribosomal proteins) show a relatively robust expression (62). Similarly, genes central in gene networks, like key pluripotency regulator *Pou5f1* (56) or encoding products that are critical to the survival of cells (also known as essential genes, since their deletion is lethal) and genes which code for multi-protein complexes have evolved to minimize their expression noise (30, 48, 54). Finally, a recent study in humans showed that long non-coding RNAs, such as anti-sense transcripts from the genomic loci corresponding to known protein-coding genes, display a higher inter-individual expression variability as compared to protein-coding genes (45) substantiating their role in adaptation.

Another biological phenomenon where the expression variability might play an important role is incomplete penetrance (71, 73). The latter study shows that in *C.elegans* mutants with more stochastic expression of *end-1* gene, a threshold for activating expression of *elt-2*, the master regulator of intestinal differentiation may or may not be reached, and hence only some of mutant embryos will develop intestine. Different levels of expression in individuals with a similar or even isogenic genetic background can explain why some individuals develop severe disease while others have a mild or even wild-type phenotype. Even individuals which are genetically identical can show phenotypic differences and even personality traits, as recently reviewed in (25). Studying transcriptomes from the

viewpoint of expression variability can provide new explanations for mechanisms of disease development.

Prerequisites for analysis of differential variability in gene expression.

Despite the high promises of differential variability analysis, several important factors should be taken into consideration when planning and performing this type of analysis.

Sufficient number of samples. While some of the studies investigating expression variability used as few as 3 samples per group (105), technical biases in library preparation and sequencing can have profound effects on the differential variability estimates. For a reproducible analysis of differential variability, a larger sample size is required in contrast to studies where a differential mean expression is tested (110). This is further exemplified below by means of power analysis in the section showcasing the differential variability analysis for mice.

Avoiding batch effects. Since technical variation can mask the effects coming from biological differences, it is important to perform all technical procedures in a single batch or, whenever that is not possible, randomly distribute samples from different groups among experiment batches.

Accounting for variability in transcript structure. While most of current studies quantify variability using number of molecules or number of sequencing reads corresponding to the gene, the structure of the transcript is rarely taken into account. Yet, variability in pre-mRNA maturation is also observed (103). At the splicing level, statistical methods were developed to identify genes with condition-specific splicing ratios (31), while variation in splicing can be defined and quantified using a recently suggested local splicing variation units (100). Future methods for differential variability analysis, therefore, should consider not only quantitative, but also structural variability of gene products.

The first two points are rather general experimental design considerations, while the latter is more specific for RNA-sequencing based profiling of gene expression.

Statistical inference of gene expression variability

Several metrics have been proposed to measure the variability of gene expression, such as variance (σ^2), the (squared) coefficient of variation (cv, also known as signal to noise ratio), Fano factor (also known as noise strength), and their robust counterparts median absolute deviation from the median (MAD), (quartile) coefficient of dispersion (COD or QCOD), etc. (74, 83, 99) (Table 1).

Applicability and interpretation of these metrics depend on how gene expression data was obtained and processed. For example, variance stabilizing transformations (VST, $f(x)$) of microarray hybridization intensities or normalized RNA counts (such as CPM - counts per million or FPKM – fragments per kilobase of transcript per million) transform mean and variance as $E[f(X)] \approx f(\mu_X)$ and $Var[f(X)] \approx (f'(\mu_X))^2 \sigma_X^2$ respectively, following the 1st-order Taylor expansion, where μ_X and σ_X^2 are original mean and variance respectively. Among commonly used VSTs are the logarithm ($log_2(X)$) and generalized logarithm ($glog_2(X) = log_2(X + \sqrt{X^2 + 1})$) functions (38). This implies that the variance of log_2 or $glog_2$ transformed variables corresponds to the squared coefficient of variation of the original variable (cv_X^2) as $Var[log_2(X)] \approx \log(2)^{-2} \frac{\sigma_X^2}{\mu_X^2} = \log(2)^{-2} cv_X^2$ and $Var[glog_2(X)] \approx \log(2)^{-2} \frac{\sigma_X^2}{\mu_X^2+1} \approx \log(2)^{-2} cv_X^2$ (for $\mu_X^2 \gg 1$). Thus, it makes no sense to estimate neither cv nor Fano factor for VST transformed variables as their variance is already proportional to cv_X^2. Similar logic applies to robust measures of variability as $MAD[log_2(X)] \approx median\left(\left|log_2(X_i/\tilde{X})\right|\right)$ and $MAD[glog_2(X)] \approx median\left(\left|log_2(X_i/\tilde{X})\right|\right)$ (for $X_i \gg 1$), and additional normalization of MAD to the median of VST transformed variable is unnecessary.

In contrast, when dealing with untransformed variables emitted by Poisson or mixed-Poisson processes (such as RNA-sequencing counts), normalization to the mean is required due to the presence of the mean-variance relationships. $Var[X] = \sigma_X^2 = \mu_X$ for Poisson and $Var[X] = \sigma_X^2 = \mu_X + \alpha_X \mu_X^2$ for mixed-Poisson distributed RNA counts, where $\alpha_X > 0$ is the overdispersion parameter (44). Then, Fano factor turns to be handy for estimation of deviation from Poisson process, as $F = \sigma_X^2/\mu_X > 1$ indicates overdispersion, while $cv_X^2 = \mu_X^{-1} + \alpha_X$ partitions noise into two asymptotically orthogonal parameters of mixed-Poisson distributions, which we refer to as Poisson and non-Poisson noise. In the section showcasing the differential variability analysis for mice we demonstrate statistical inference of both μ_X and α_X parameters for genes' RNA counts.

So far, statistical inference of expression variability is limited to only a few tools. For instance, tools, such as AEGS and pathVar aim to discover biological pathways, for which the expression variability changes. AEGS is a webserver that uses case-control data and allows to identify which of pre-defined gene sets (e.g. genes belonging to the same gene ontology

category) are more variable expressed and ranks variability of individual genes within each set (32). The tool is also available as standalone program and can, in principle, be easily integrated into RNA-Seq analysis pipelines. PathVar enables case-control pathway-based interpretation of gene expression variability, but can also compare a single group of samples against a background distribution (99). This tool is available from Bioconductor collection of packages, provides a broad choice of variability measures and can also become part of routine transcriptome analysis.

Another tool, MDseq employs a generalized linear model (GLM) to estimate statistically significant changes in both expression mean and variability in response to experimental factors (74). Although MDseq considerably expands the standard GLM approach employed in many tools for differential gene expression analysis, its current implementation seems to be limited to a fixed effect negative binomial (NB) model (74). To that, MDseq parametrization of the NB implies a linear mean-variance relationship for RNA counts: $Var(X) = \mu\varphi$, while many RNA-seq studies suggest a quadratic relationship (58). In fact, a quadratic mean-variance relationship originates from the mixed-Poisson nature of a stochastic process driving the RNA synthesis and degradation (21, 40, 66, 72).

In brief, for a mixed-Poisson processes, the Poisson rate (μ), represented by a ratio of RNA synthesis to degradation rates, is assumed to be a random variable with expectation $E(\mu) = \mu$ and the variance defined by an underlying mixing distribution $g_\mu(\mu)$. The mixed Poisson distribution of RNA counts takes the following general form: $P(X = x) = \int_0^\infty \frac{e^{-\mu}\mu^x}{x!} g_\mu(\mu)\, d\mu$, where mixing distribution $g_\mu(\mu)$ can take on any parametric form depending on upstream cellular drives (21). For example, promoter switching between active and inactive states (bursts) leads under limiting conditions to a gamma distribution of the Poisson rate (μ). As a result, the cell-to-cell distribution of the RNA copy number follows a gamma-Poisson distribution (also known as a negative binomial, NB) (21, 72). Likewise, the NB distribution empirically fits well to RNA sequencing counts from both tissues and cell populations (58).

For any mixed-Poisson process, i.e. independent of a specific form of the $g_\mu(\mu)$, a total variance and noise (a squared coefficient of variation of RNA counts) sums from the Poisson (1st summand) and non-Poisson (2nd summand) parts as: $Var[X] = \mu + \alpha\mu^2$, $cv^2(X) = \mu^{-1} + \alpha$ respectively (44, 79). Non-Poisson variation – α is often referred to as the

overdispersion parameter or the biological coefficient of variation ($\alpha = bcv^2$) (58). The Poisson and non-Poisson variation are also assigned as "intrinsic" and "extrinsic" respectively (68). Thus, the goal of differential gene expression analysis is to estimate the average RNA copy number - μ, while that of differential gene noise analysis is to estimate overdispersion - α from a distribution of RNA counts.

A showcase for differential gene expression variability analysis using GAMLSS

Here we propose to utilize GAMLSS to assess the effects of biological factors on a gene's Poisson (μ^{-1}) and non-Poisson (α) variation. GAMLSS stands for generalized additive model for location, scale and shape and offers immense flexibility for semi-parametric mixed effect modelling of up to four distribution parameters (78, 91).

The suggested analysis strategy has several advantages. First, GAMLSS comes with an extensive list of mixed-Poisson distributions along with their zero inflated/adjusted variants (79). Second, GAMLSS allows for the fitting of mixed effect models to RNA counts. And third, smoothing terms (splines) can also be used to model non-linear relations of mixed-Poisson distribution parameters with continuous experimental variables such as age. These factors combined give it a much better control in the modelling of differential gene expression and variability.

To demonstrate GAMLSS at work, we provide a brief re-analysis of age-dependent changes in the overdispersion (non-Poisson variation) for genes expressed in liver samples taken from young and old C57BL/6J mice (61). All computer programs used here and description of the analysis are available as GitHub repository (https://github.com/Vityay/ExpVarQuant).

We modeled genes' RNA counts using the $NB(\mu, \alpha)$ distribution parametrized with respect to the mean (μ) and non-Poisson variation (α) in such a way that the quadratic mean-variance relationship holds. The probability mass function for independent and identically distributed RNA counts (X) for a given gene: $X \overset{\text{ind}}{\sim} NB(\mu, \alpha)$ is defined as:

$$P(X = x) = \frac{\Gamma\left(\frac{1}{\alpha}+x\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(x+1)}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{x},$$

with expectation (mean) and variance of RNA counts: $E[X] = \mu$, $Var[X] = \mu + \alpha\mu^2$, and $cv^2(X) = \mu^{-1} + \alpha$.

Then, we specified a GAMLSS model to account for the age (young – 5 months and old – 20 months old mice) effect on both the mean RNA counts and the overdispersion:

$$\log(X_i) \sim age_j \beta_{\mu_j} + log(N_i),$$

$$\log(\alpha) \sim age_j \beta_{\alpha_j},$$

where $i = 1, \dots, n$ is i^{th} observation of gene's mRNA counts (X_i); $j = 1, \dots, p$ is j^{th} factor level (young – 5 weeks, old – 20 weeks); and $log(N_i)$ is offset vector represented by library sizes. The first equation of GAMLSS specifies a model of a factor effect, namely age_j, on library size (N_i) normalized mean mRNA counts ($\mu_j = e^{\beta_{\mu_j}}$, $cpm_j = 10^6 \mu_j$). Essentially, this part of the model corresponds to a GLM model of differential gene expression (58), however, GAMLSS allows for more flexibility as random effects and smoothing terms can also be included (91). The second equation of GAMLSS models a factor effect on non-Poisson noise (α), where β_{α_j} is a maximum-likelihood estimation of overdispersion parameter ($\alpha_j = e^{\beta_{\alpha_j}}$).

Significance values of age-mediated changes in μ and α parameters of the $NB(\mu, \alpha)$ were assessed for each gene with likelihood ratio tests (LR). For a given gene, LR test statistic for changes in mean RNA counts between old and young mice was calculated as following:

$$D_\mu = -2log \frac{\text{likelihood for reduced model}}{\text{likelihood for GAMLSS model}} = -2log \frac{\mathcal{L}(\mu_0, \alpha_j \mid X_i)}{\mathcal{L}(\mu_j, \alpha_j \mid X_i)},$$

Where the reduced model omits factor effect (age) from the model of μ: $\log(X_i) \sim \beta_{\mu_0} + log(N_i)$, while the age effect on non-Poisson noise was still accounted for. It can be readily noted that the estimation of differential gene expression by GAMLSS differs from that by classical GLM as the latter estimates only the shared overdispersion (58). In brief, the GLM model is specified as:

$$\log(X_i) \sim age_j \beta_{\mu_j} + log(N_i),$$

$$\log(\alpha) \sim \beta_{\alpha_0}$$

in GAMLSS notation and the LR test statistic is calculated as:

$$D_{\mu_{GLM}} = -2log \frac{\text{likelihood for null model}}{\text{likelihood for GLM model}} = -2log \frac{\mathcal{L}(\mu_0, \alpha_0 \mid X_i)}{\mathcal{L}(\mu_j, \alpha_0 \mid X_i)},$$

where null model omits factor effect on both μ and α. Finally, LR test statistic for changes in non-Poisson noise was calculated by comparing GLM model (as reduced model for α) with full GAMLSS model:

$$D_\alpha = -2log \frac{\text{likelihood for GLM model}}{\text{likelihood for GAMLSS model}} = -2log \frac{\mathcal{L}(\mu_j, \alpha_0 \mid X_i)}{\mathcal{L}(\mu_j, \alpha_j \mid X_i)}.$$

460 D_μ, $D_{\mu_{GLM}}$ and D_α are asymptotically χ^2-distributed with degrees of freedom equal to a
461 difference between the number of compared models' parameters. Thus, from this example
462 it is clear that GAMLSS is an extension of a GLM model allowing for the estimation of factor
463 effects on both parameters of the distribution of RNA counts, namely mean and
464 overdispersion (non-Poisson noise).

465 We excluded genes with zero counts in any of the samples from the analysis as this
466 might bias the estimation of non-Poisson variation. In fact, an excess of zeros in RNA-seq
467 data imposes a certain problem for statistical inference of the distribution parameters for
468 RNA counts. Indeed, in many cases it is impossible to discriminate whether observing a zero
469 is the result of a gene being silenced or whether it is observed due to an insufficient
470 sequencing depth causing dropouts of the lowly expressed genes. In principle, the former
471 case corresponds to a zero-adjusted model, while the latter – to a zero-inflated model, and
472 both could be fitted by GAMLSS. However, neither of these assumptions alone resolves the
473 uncertainty that zero values introduce to transcriptome analysis.

474 Having estimated the parameters μ and α for liver genes expressed in young and old
475 mice, we noted that their absolute values were practically uncorrelated ($\rho(\mu, \alpha) \to 0$). This
476 could be attributed directly to the given parametrization of the $NB(\mu, \alpha)$, which implies an
477 asymptotic independence of the estimated parameters. It follows from the Fisher
478 information matrix as its element $I_{\mu\alpha} = -E\frac{\partial^2}{\partial\mu\partial\alpha}\log\big(P(X|\,\mu,\alpha)\big) = 0$. To that, changes in
479 the mean gene expression and the non-Poisson variation occurring with age were also
480 almost uncorrelated ($\rho(\Delta\mu, \Delta\alpha) \to 0$). Testing under the assumption that the cellular RNA
481 concentration (total number of RNA molecules per cell) is the same for the samples taken
482 from young and old mice, we scored about a comparable number of genes for which the
483 mean RNA counts either increased or decreased significantly with age (Fig. 2A, 3A).
484 Estimation of the mean also yielded the estimation of the Poisson variation as they are
485 reciprocal to each other (Poisson variation = μ^{-1}). In contrast to the Poisson variation, non-
486 Poisson variation increased with age (Fig. 2B). Importantly, applying the GAMLSS model
487 enabled for the identification of genes for which the non-Poisson variation, but not the
488 mean, changed significantly with age (Fig. 2B, 3B).

489 However, it must be noted that the relative standard errors of overdispersion estimates
490 tend to be larger than that of mean estimates. As a result, this lowers the statistical power

of likelihood ratio test for factor effects on non-Poisson variation. This is evident from the power analysis of LR tests for fold changes in mean and overdispersion (Fig. 2C, D). Although a derivation of the analytical form for the power of LR tests for complex models deems impossible, this can be circumvented by a simulation method. To this end, a thousand pairs of samples of NB distributed random variables were generated with the given parameters μ_0 (counts) and α_0 (non-Poisson noise) for reference samples and fold changes (δ) in one of the NB parameters for test samples. Then, LR tests were applied comparing simulated reference samples $NB(\mu_0, \alpha_0)$ with test samples $NB(\delta\mu_0, \alpha_0)$ and $NB(\mu_0, \delta\alpha_0)$. The power of LR tests for $\mu_0 \neq \delta\mu_0$ (Fig. 2C) and $\alpha_0 \neq \delta\alpha_0$ (Fig. 2D) was then estimated as proportion of true positives at significance level of < 0.05. Obviously for all tested configurations of NB (μ_0: $\{10, 100, 1000\}$ and α_0: $\{0.1, 0.25, 0.5\}$) the power of LR tests for mean and overdispersion increased with an increasing sample size. To that, the power of LR tests for fold changes in mean counts (Fig. 2C) is higher than that of non-Poisson noise (Fig. 2D). Unexpectedly though, the power of LR tests tends to increase, especially for the tests comparing overdispersion, with increasing μ_0 irrespectively of the presence or absence of an offset parameter, which simulates library size. This suggests that an increase of sample size and sequencing depth (library size) will eventually increase the statistical power of tests aimed at comparing changes in mean expression and non-Poisson noise.

The expression variability analysis provides additional insights into dataset

To identify biological pathways associated with the age-mediated increase in non-Poisson variation, we fitted a ridge regression model to the $\log_2$ fold change in overdispersion using KEGG annotations of genes as a model matrix (Fig. 4A) (35, 43). Such an approach circumvents the problem of pathways overrepresentation analysis associated with the necessity to select a threshold for statistical significance. It is also well suited for the analysis of non-Poisson variation when a common trend for genes is to increase in variability with age. As a result, the KEGG-pathway ridge regression model revealed several pathways, such as the complement and coagulation cascades, amino acid (Val, Leu, Ile) degradation, chemokine signaling and others for which non-Poisson variation increased in aged mice (Fig. 3B, 4B).

Fluctuation-response relationship for RNA counts

Gene expression noise is thought to drive gene expression plasticity due to a fluctuation-response relationship (49, 84). This implies that an absolute change in the

523 expectation (μ) of some measurable quantity (X) in response to an influence is proportional
524 to its initial variance: $|\mu_1 - \mu_0| \sim Var(X)$. However, this relationship holds only true for
525 Gaussian-like distributed quantities under the assumption of a fixed variance:
526 $Var(X_1) \sim Var(X_0)$. Nonetheless, if log transformed RNA counts approximate a Gaussian-
527 like distribution, then the fluctuation-response relationship takes on the following form:
528 $|\log(\mu_1/\mu_0)| \sim \alpha = bcv^2$, as a result of the Taylor expansions for the moments for genes
529 expressed at large copy number ($\mu \gg 1$). We noted a modest, but significant, positive
530 correlation between absolute $\log_2$ fold changes in the mean gene expression for old and
531 young mice with non-Poisson variation for young mice (Fig. 5A). A lack of a stronger
532 correlation could be due to the violation of the fluctuation-response assumption of a fixed
533 variance or overdispersion for log-transformed variables. In general, this substantiates the
534 fluctuation-response relationship for the RNA copy number.
535
536 **Estimates of gene variation from tissues retain information on gene state**
537 **determinants of non-Poisson noise.**
538 Finally, we wondered if the estimate of non-Poisson variation from RNA-sequencing
539 data of cell populations contain information on gene state determinants. To this end, we
540 compared the genes' non-Poisson variation estimates with their promoter DNA-sequence
541 composition. First, we noted that on average, that the non-Poisson variation was higher for
542 genes that were regulated by TATA-containing promoters (Fig. 5B). Second, in accordance
543 with the fluctuation-response relationship (Fig. 5A), aging induced more pronounced
544 changes in the mean expression of genes with TATA-containing promoters (Fig. 5B, C).
545 Overall, this result is in agreement with the TATA-mediated promoter fluctuation caused by
546 a competition between distinct TBP-co-activator complexes (77, 82, 87) and it substantiates
547 that gene state signals are retained in cell population estimates of non-Poisson variation.
548 To conclude this brief showcase of GAMLSS, we advocate for the use of this framework
549 to dissect the determinants of both the mean RNA counts and the non-Poisson variation as
550 two independent parameters of gene expression network.
551 **Combining other -omics data with RNA-seq can lead to new discoveries.**
552 A connection between the gene expression variability measured on different levels:
553 cell-to-cell, inter-individual and inter-population has been suggested previously (23, 25). The
554 rapid development of accessible and cost-efficient methods for single-cell RNA-seq (scRNA-

555 seq) will provide us with improved estimates of cell-to-cell variability in gene expression
556 (70). Flow cytometry techniques can help in the further separation into (so called / the
557 suggested) macro-heterogeneity, which is the variability that encompasses both the on- and
558 off- state of genes, as well as the micro-heterogeneity, which represents the variability in
559 gene expression of genes in different (37). Further, recently generated large transcriptome
560 datasets for hundreds of individuals (2, 47) should increase our understanding of
561 transcriptome variability at population level.

562 Apart from transcriptomics data, large sets of epigenetics data will be of great value.
563 For example, the changing landscape of histone modifications with age has been established
564 (89), as has the property of histone modifications to be associated with the average gene
565 expression and variation in gene expression (108). Similarly, the beneficial effects of
566 alterations in diet have been shown to extend the lifespan of mice (7), as has the
567 methylation of genes and the consequent variation in expression been shown to contribute
568 to the pathophysiology of mice on a high fat diet (113). In line with these two observations,
569 it has been shown that the suppression of inter-individual variation has positive effects on
570 the lifespans of *C. elegans* (75).

571 Finally, when speaking of gene expression variability, it is important to consider how the
572 variability in RNA copy number translates to variability at a protein level. Often there seems
573 to be a discrepancy between the amount of RNA transcribed and the amount of the
574 matching protein being produced within samples (64). Yet, many principles of gene noise
575 have been derived by quantifying reporter genes expression on protein level, such as two-
576 color reporter assay (26, 94). To that, derivations of protein fluctuations from theoretical
577 models of stochastic gene expression highlight the contribution of RNA-level noise to
578 protein-level noise (68). Thus, it makes it reasonable to propose that gene expression
579 variability might propagate from RNA to protein, from protein to cell, from cell to tissue and
580 from tissue to organism.

581 To conclude, the analysis of differential transcriptome variability complements the
582 standard analysis of differential gene expression and reveals another dimension of
583 expression analysis. With the further development of tools and with a wider acceptance of
584 these methods, we will advance our understanding of the mechanisms underlying the
585 regulation of transcription, common physiological traits and disease predispositions.

Table 1. Commonly used measures of variability

Coefficient of variation (signal to noise ratio)	$cv = \sigma/\mu$
Fano factor (noise strength)	$F = \sigma^2/\mu$
Median Absolute Deviation from the Median	$MAD = median(\lvert X_i - \tilde{X} \rvert)$
Coefficient of Dispersion	$COD = MAD/\tilde{X}$
Quartile Coefficient of Dispersion	$QCOD = (Q_3 - Q_1)/(Q_3 + Q_1)$

$\tilde{X}$ - median; Q_1 and Q_3 are the 1st and 3rd quartiles respectively.

Figure legends

Figure 1. A model depicting factors influencing the gene expression variability/noise. Key equations depicting the partitioning of variance and squared coefficient of variations into Poisson (blue, Pois.) and non-Poisson (red, non-Pois.) variability/noise are shown. Such partitioning holds true for any mixed-Poisson distribution, where the Poisson rate μ is a random variable distributed with expectation $\langle\mu\rangle$ and variance $Var[\mu]$. Key equations for the expectation ($E[RNA]$), variance ($Var[RNA]$) and noise (cv^2(RNA)) for two-state promoter model are expressed in terms of burst size (b) and burst frequency (f_b). See text for further details.

Figure 2. A GAMLSS analysis of age-mediated changes in gene expression and non-Poisson noise.

A) Boxplots of a GAMLSS estimations of the mean mRNAs copy numbers (counts per million mapped reads, cpm) for genes expressed in the liver of young (5 months, n = 6) and old (20 months, n = 6) C57BL/6J mice (left panel). Scatter plot of genes' mean mRNA copy number in young and old mice (middle panel) and a boxplot of $\log_2$ fold changes in expression between old and young mice (right panel). Significantly up- and down-regulated genes (false discovery rate, FDR ≤ 0.05) are indicated in red and blue respectively. In boxplots, the box spans the interquartile range (IQR) from 25% (Q1) to 75% (Q3) and the middle line indicates 50% (median). Whiskers span to 1.5 IQR from the lower (Q1) and upper (Q3) quartiles or are truncated to the min or max values, if those are within 1.5 IQR.

B) GAMLSS estimation of non-Poisson variability in mRNAs copy numbers (left panel). A scatter plot of genes' estimates of non-Poisson variability in young and old mice (middle

612 panel) and a boxplot of $\log_2$ fold changes in non-Poisson variability with age (right panel).
613 Genes for which the non-Poisson noise increased or decreased significantly with age are
614 marked in red or blue respectively.
615 **C)** Heatmap depicting a power analysis of the likelihood ratio (LR) test for fold changes (δ) in
616 μ_0 (mean counts). For each power analysis (1000) pairs of samples from reference
617 $NB(\mu_0, \alpha_0)$ and test $NB(\delta\mu_0, \alpha_0)$ distributions were simulated with $\mu_0 \in \{10,100,1000\}$,
618 $\alpha_0 \in \{0.1,0.25,0.5\}$ and $\delta \in \left\{\frac{1}{4}, \frac{1}{3}, \ldots, 3, 4\right\}$. Sample sizes were $\{5,10, \ldots, 100,1000\}$. Null
619 hypothesis: $H_0: \mu_0 = \delta\mu_0$ were rejected at significance level of 0.05 and power was
620 calculated as the probability of rejecting H_0. Red indicates high power, white -low.
621 **D)** Heatmap depicting a power analysis of the likelihood ratio (LR) test for fold change (δ) in
622 α_0 (non-Poisson noise).
623
624 **Figure 3.** Examples of differentially expressed genes (**A**) and genes showing increase in non-
625 Poisson variability with age (**B**). Upper panel, boxplots of selected liver genes' mRNAs copy
626 numbers (expressed as $\log_2$(cpm)) for young (green, n = 6) and old (red, n = 6) C57BL/6J
627 mice. Whiskers extend to minimum and maximum values. Middle panel, boxplots of
628 $\log_2$(cpm) residual values corrected for genes' grand mean expression for young and old
629 mice (~ gene). Lower panel, boxplots of $\log_2$(cpm) residuals corrected for genes' group-wise
630 mean expression in young and old mice (~gene:age). The middle panel serves to illustrate
631 differential gene expression, while the lower panel shows whether the gene expression
632 variability is affected by age. Genes were selected based on significance of the age-
633 mediated changes in mean mRNA counts (**A**, $FDR_{cpm} \leq 0.05$) or changes in non-Poisson
634 variability (**B**, $FDR_{\text{non-Pois. variability}} \leq 0.05$). For (**B**), note an increase in $\log_2$(cpm) variability for
635 selected genes in population of 20 weeks old mice due to an increase in non-Poisson
636 variability with age as compared to 5 weeks mice. Left panel in (**B**) shows genes associated
637 with complement and coagulation cascades according to KEGG annotation, the right panel
638 shows a selection of 30 genes with the highest statistically significant gain in non-Poisson
639 variability.
640
641 **Figure 4.** Pathway analysis of age-mediated changes in non-Poisson variability.

A) Ridge regression model predicting age-mediated changes in non-Poisson variability based on the genes' KEGG pathway annotations.
B) Top 20 KEGG pathways associated with age-mediated increase in non-Poisson variability. Pathways were selected based on the ranking of model coefficients.

Figure 5. A) Relationships between the initial non-Poisson variability in young mice and the age-mediated responses in the mean mRNAs counts. Gene expression responses are represented as absolute $\log_2$ ratios (top panel) of mean mRNA counts in old and young mice. GAMLSS estimates of genes' non-Poisson variability in young mice are given as ranked values ranging from lowest (1) to highest (10). Spearman correlation coefficients are shown. Trend lines were generated by LOESS local regression.
B-C) TATA-box associated with increased non-Poisson variability and age-mediated response in mean expression levels. **(B)** Boxplots show the initial non-Poisson variability in 5 months old mice (young, upper panel) and absolute changes in the mean gene expression (lower panel) for mouse genes classified according to all possible combinations of four promoter motifs: the TATA-box, Initiator (Inr), CCAAT-box and GC-box. A group of genes lacking any of those is labelled as "none". **(C)** Scatterplot of genes' group-wise medians in the initial non-Poisson variability at age of 5 months and the absolute changes in mean gene expression levels between old and young mice. Genes containing a TATA-box in any of these combinations in their promoters tend to have a higher non-Poisson variability and respond stronger to age with respect to the changes in mean expression levels. The Pearson correlation coefficient and significance are indicated.

References

1. **Adelman K**, **Henriques T**. Transcriptional speed bumps revealed in high resolution. *Nature* 560: 560–561, 2018.
2. **Aguet F**, **Brown AA**, **Castel SE**, **Davis JR**, **He Y**, **Jo B**, **Mohammadi P**, **Park Y**, **Parsana P**, **Segrè A V.**, **Strober BJ**, **Zappala Z**, **Cummings BB**, **Gelfand ET**, **Hadley K**, **Huang KH**, **Lek M**, **Li X**, **Nedzel JL**, **Nguyen DY**, **Noble MS**, **Sullivan TJ**, **Tukiainen T**, **MacArthur DG**, **Getz G**, **Addington A**, **Guan P**, **Koester S**, **Little AR**, **Lockhart NC**, **Moore HM**, **Rao A**, **Struewing JP**, **Volpi S**, **Brigham LE**, **Hasz R**, **Hunter M**, **Johns C**, **Johnson M**, **Kopen G**, **Leinweber WF**, **Lonsdale JT**, **McDonald A**, **Mestichelli B**, **Myer K**, **Roe B**, **Salvatore M**, **Shad S**, **Thomas JA**, **Walters G**, **Washington M**, **Wheeler J**, **Bridge J**, **Foster BA**, **Gillard BM**, **Karasik E**, **Kumar R**, **Miklos M**, **Moser MT**, **Jewell SD**, **Montroy RG**, **Rohrer DC**, **Valley D**, **Mash DC**, **Davis DA**, **Sobin L**, **Barcus ME**, **Branton PA**, **Abell NS**, **Balliu B**, **Delaneau O**, **Frésard L**, **Gamazon ER**, **Garrido-Martín D**, **Gewirtz ADH**, **Gliner G**, **Gloudemans MJ**, **Han B**, **He AZ**, **Hormozdiari F**, **Li X**, **Liu B**, **Kang EY**, **McDowell IC**, **Ongen H**, **Palowitch JJ**, **Peterson CB**, **Quon G**, **Ripke S**, **Saha A**, **Shabalin AA**, **Shimko TC**, **Sul JH**, **Teran NA**, **Tsang EK**, **Zhang H**, **Zhou Y-H**, **Bustamante CD**, **Cox NJ**, **Guigó R**, **Kellis M**, **McCarthy MI**, **Conrad DF**, **Eskin E**, **Li G**, **Nobel AB**, **Sabatti C**, **Stranger BE**, **Wen X**, **Wright FA**, **Ardlie KG**, **Dermitzakis ET**, **Lappalainen T**, **Aguet F**, **Ardlie KG**, **Cummings BB**, **Gelfand ET**, **Getz G**, **Hadley K**, **Handsaker RE**, **Huang KH**, **Kashin S**, **Karczewski KJ**, **Lek M**, **Li X**, **MacArthur DG**, **Nedzel JL**, **Nguyen DT**, **Noble MS**, **Segrè A V.**, **Trowbridge CA**, **Tukiainen T**, **Abell NS**, **Balliu B**, **Barshir R**, **Basha O**, **Battle A**, **Bogu GK**, **Brown A**, **Brown CD**, **Castel SE**, **Chen LS**, **Chiang C**, **Conrad DF**, **Cox NJ**, **Damani FN**, **Davis JR**, **Delaneau O**, **Dermitzakis ET**, **Engelhardt BE**, **Eskin E**, **Ferreira PG**, **Frésard L**, **Gamazon ER**, **Garrido-Martín D**, **Gewirtz ADH**, **Gliner G**, **Gloudemans MJ**, **Guigo R**, **Hall IM**, **Han B**, **He Y**, **Hormozdiari F**, **Howald C**, **Kyung Im H**, **Jo B**, **Yong Kang E**, **Kim Y**, **Kim-Hellmuth S**, **Lappalainen T**, **Li G**, **Li X**, **Liu B**, **Mangul S**, **McCarthy MI**, **McDowell IC**, **Mohammadi P**, **Monlong J**, **Montgomery SB**, **Muñoz-Aguirre M**, **Ndungu AW**, **Nicolae DL**, **Nobel AB**, **Oliva M**, **Ongen H**, **Palowitch JJ**, **Panousis N**, **Papasaikas P**, **Park Y**, **Parsana P**, **Payne AJ**, **Peterson CB**, **Quan J**, **Reverter F**, **Sabatti C**, **Saha A**, **Sammeth M**, **Scott AJ**, **Shabalin AA**, **Sodaei R**, **Stephens M**, **Stranger BE**, **Strober BJ**, **Sul JH**, **Tsang EK**, **Urbut S**, **van de Bunt M**, **Wang G**, **Wen X**, **Wright FA**, **Xi HS**, **Yeger-Lotem E**, **Zappala Z**, **Zaugg JB**, **Zhou Y-H**, **Akey JM**, **Bates D**, **Chan J**, **Chen LS**, **Claussnitzer M**, **Demanelis K**, **Diegel M**, **Doherty JA**, **Feinberg AP**, **Fernando MS**, **Halow J**, **Hansen KD**, **Haugen E**, **Hickey PF**, **Hou L**, **Jasmine F**, **Jian R**, **Jiang L**, **Johnson A**, **Kaul R**, **Kellis M**, **Kibriya MG**, **Lee K**, **Billy Li J**, **Li Q**, **Li X**, **Lin J**, **Lin S**, **Linder S**, **Linke C**, **Liu Y**, **Maurano MT**, **Molinie B**, **Montgomery SB**, **Nelson J**, **Neri FJ**, **Oliva M**, **Park Y**, **Pierce BL**, **Rinaldi NJ**, **Rizzardi LF**, **Sandstrom R**, **Skol A**, **Smith KS**, **Snyder MP**, **Stamatoyannopoulos J**, **Stranger BE**, **Tang H**, **Tsang EK**, **Wang L**, **Wang M**, **Van Wittenberghe N**, **Wu F**, **Zhang R**, **Nierras CR**, **Branton PA**, **Carithers LJ**, **Guan P**, **Moore HM**, **Rao A**, **Vaught JB**, **Gould SE**, **Lockart NC**, **Martin C**, **Struewing JP**, **Volpi S**, **Addington AM**, **Koester SE**, **Little AR**, **Brigham LE**, **Hasz R**, **Hunter M**, **Johns C**, **Johnson M**, **Kopen G**, **Leinweber WF**, **Lonsdale JT**, **McDonald A**, **Mestichelli B**, **Myer K**, **Roe B**, **Salvatore M**, **Shad S**, **Thomas JA**, **Walters G**, **Washington M**, **Wheeler J**, **Bridge J**, **Foster BA**, **Gillard BM**, **Karasik E**, **Kumar R**, **Miklos M**, **Moser MT**, **Jewell SD**, **Montroy RG**,

Rohrer DC, **Valley DR**, **Davis DA**, **Mash DC**, **Undale AH**, **Smith AM**, **Tabor DE**, **Roche N V.**, **McLean JA**, **Vatanian N**, **Robinson KL**, **Sobin L**, **Barcus ME**, **Valentino KM**, **Qi L**, **Hunter S**, **Hariharan P**, **Singh S**, **Um KS**, **Matose T**, **Tomaszewski MM**, **Barker LK**, **Mosavel M**, **Siminoff LA**, **Traino HM**, **Flicek P**, **Juettemann T**, **Ruffier M**, **Sheppard D**, **Taylor K**, **Trevanion SJ**, **Zerbino DR**, **Craft B**, **Goldman M**, **Haeussler M**, **Kent WJ**, **Lee CM**, **Paten B**, **Rosenbloom KR**, **Vivian J**, **Zhu J**. Genetic effects on gene expression across human tissues. *Nature* 550: 204–213, 2017.

3. **Alexander RD**, **Innocente SA**, **David Barrass J**, **Beggs JD**. Splicing-Dependent RNA Polymerase Pausing in Yeast. *Mol Cell* 40: 582–593, 2010.
4. **Allis CD**, **Jenuwein T**. The molecular hallmarks of epigenetic control. *Nat Rev Genet* 17: 487, 2016.
5. **Azzi A**, **Dallmann R**, **Casserly A**, **Rehrauer H**, **Patrignani A**, **Maier B**, **Kramer A**, **Brown SA**. Circadian behavior is light-reprogrammed by plastic DNA methylation. *Nat Neurosci* 17: 377–82, 2014.
6. **Bahar R**, **Hartmann CH**, **Rodriguez KA**, **Denny AD**, **Busuttil RA**, **Dollé MET**, **Calder RB**, **Chisholm GB**, **Pollock BH**, **Klein CA**, **Vijg J**. Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* 441: 1011–1014, 2006.
7. **Barrington WT**, **Wulfridge P**, **Wells AE**, **Rojas CM**, **Howe SYF**, **Perry A**, **Hua K**, **Pellizzon MA**, **Hansen KD**, **Voy BH**, **Bennett BJ**, **Pomp D**, **Feinberg AP**, **Threadgill DW**. Improving Metabolic Health Through Precision Dietetics in Mice. *Genetics* 208: 399–417, 2018.
8. **Bauer T**, **Trump S**, **Ishaque N**, **Thürmann L**, **Gu L**, **Bauer M**, **Bieg M**, **Gu Z**, **Weichenhan D**, **Mallm J-P**, **Röder S**, **Herberth G**, **Takada E**, **Mücke O**, **Winter M**, **Junge KM**, **Grützmann K**, **Rolle-Kampczyk U**, **Wang Q**, **Lawerenz C**, **Borte M**, **Polte T**, **Schlesner M**, **Schanne M**, **Wiemann S**, **Geörg C**, **Stunnenberg HG**, **Plass C**, **Rippe K**, **Mizuguchi J**, **Herrmann C**, **Eils R**, **Lehmann I**. Environment-induced epigenetic reprogramming in genomic regulatory elements in smoking mothers and their children. *Mol Syst Biol* 12: 861, 2016.
9. **Bird AP**. Gene number, noise reduction and biological complexity. *Trends Genet* 11: 94–100, 1995.
10. **Blackledge NP**, **Klose RJ**. CpG island chromatin: A platform for gene regulation. *Epigenetics* 6: 147–152, 2011.
11. **Blake WJ**, **Balázsi G**, **Kohanski MA**, **Isaacs FJ**, **Murphy KF**, **Kuang Y**, **Cantor CR**, **Walt DR**, **Collins JJ**. Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Mol Cell* 24: 853–865, 2006.
12. **Bódi Z**, **Farkas Z**, **Nevozhay D**, **Kalapis D**, **Lázár V**, **Csörgő B**, **Nyerges Á**, **Szamecz B**, **Fekete G**, **Papp B**, **Araújo H**, **Oliveira JL**, **Moura G**, **Santos MAS**, **Székely Jr T**, **Balázsi G**, **Pál C**. Phenotypic heterogeneity promotes adaptive evolution. *PLOS Biol* 15: e2000644, 2017.
13. **Braun KA**, **Young ET**. Coupling mRNA Synthesis and Decay Downloaded from. *Mol Cell Biol* 34: 4078–4087, 2014.
14. **Brown CR**, **Mao C**, **Falkovskaia E**, **Jurica MS**, **Boeger H**. Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. *PLoS Biol* 11: e1001621, 2013.
15. **Cellerino A**, **Ori A**. What have we learned on aging from omics studies? *Semin Cell Dev Biol* 70: 177–189, 2017.
16. **Chathoth KT**, **David Barrass J**, **Webb S**, **Beggs JD**. A Splicing-Dependent

Transcriptional Checkpoint Associated with Prespliceosome Formation. *Mol Cell* 53: 779–790, 2014.

17. **Chereji R V**, **Kan T-W**, **Grudniewska MK**, **Romashchenko A V**, **Berezikov E**, **Zhimulev IF**, **Guryev V**, **Morozov A V**, **Moshkin YM**. Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Res* 44: 1036–51, 2016.
18. **Choi JK**, **Kim Y-J**. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat Genet* 41: 498–503, 2009.
19. **Churchman LS**, **Weissman JS**. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469: 368–73, 2011.
20. **Collings CK**, **Anderson JN**. Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics Chromatin* 10: 18, 2017.
21. **Dattani J**, **Barahona M**. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *J R Soc Interface* 14: 20160833, 2017.
22. **Dey SS**, **Foley JE**, **Limsirichai P**, **Schaffer D V**, **Arkin AP**. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol Syst Biol* 11: 806, 2015.
23. **Dong D**, **Shao X**, **Deng N**, **Zhang Z**. Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res* 39: 403–413, 2011.
24. **Ecker S**, **Chen L**, **Pancaldi V**, **Bagger FO**, **Fernández JM**, **Carrillo de Santa Pau E**, **Juan D**, **Mann AL**, **Watt S**, **Casale FP**, **Sidiropoulos N**, **Rapin N**, **Merkel A**, **BLUEPRINT Consortium HG**, **Stunnenberg HG**, **Stegle O**, **Frontini M**, **Downes K**, **Pastinen T**, **Kuijpers TW**, **Rico D**, **Valencia A**, **Beck S**, **Soranzo N**, **Paul DS**. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol* 18: 18, 2017.
25. **Ecker S**, **Pancaldi V**, **Valencia A**, **Beck S**, **Paul DS**. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. *BioEssays* 40: 1700148, 2018.
26. **Elowitz MB**, **Levine AJ**, **Siggia ED**, **Swain PS**. Stochastic Gene Expression in a Single Cell. *Science* 297: 1183–1186, 2002.
27. **Faure AJ**, **Schmiedel JM**, **Lehner B**. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst* 5: 471–484.e4, 2017.
28. **Feinberg AP**, **Irizarry RA**. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* 107 Suppl 1: 1757–64, 2010.
29. **Fraga MF**, **Ballestar E**, **Paz MF**, **Ropero S**, **Setien F**, **Ballestar ML**, **Heine-Suñer D**, **Cigudosa JC**, **Urioste M**, **Benitez J**, **Boix-Chornet M**, **Sanchez-Aguilera A**, **Ling C**, **Carlsson E**, **Poulsen P**, **Vaag A**, **Stephan Z**, **Spector TD**, **Wu Y-Z**, **Plass C**, **Esteller M**. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 102: 10604–9, 2005.
30. **Fraser HB**, **Hirsh AE**, **Giaever G**, **Kumm J**, **Eisen MB**. Noise Minimization in Eukaryotic Gene Expression. *PLoS Biol* 2: e137, 2004.
31. **Gonzalez-Porta M**, **Calvo M**, **Sammeth M**, **Guigo R**. Estimation of alternative splicing variability in human populations. *Genome Res* 22: 528–538, 2012.
32. **Guan J**, **Chen M**, **Ye C**, **Cai JJ**, **Ji G**. AEGS: identifying aberrantly expressed gene sets for differential variability analysis. *Bioinformatics* 34: 881–883, 2018.
33. **Haimovich G**, **Choder M**, **Singer RH**, **Trcek T**. The fate of the messenger is pre-determined: a new model for regulation of gene expression. *Biochim Biophys Acta* 1829: 643–53, 2013.

34. **Hannum G**, **Guinney J**, **Zhao L**, **Zhang L**, **Hughes G**, **Sadda S**, **Klotzle B**, **Bibikova M**, **Fan J-B**, **Gao Y**, **Deconde R**, **Chen M**, **Rajapakse I**, **Friend S**, **Ideker T**, **Zhang K**. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 49: 359–367, 2013.
35. **Hastie T**, **Tibshirani R**, **Friedman J**. Basis Expansions and Regularization. In: *The Elements of Statistical Learning*. Springer, p. 139–189, 2009.
36. **Hornung G**, **Bar-Ziv R**, **Rosin D**, **Tokuriki N**, **Tawfik DS**, **Oren M**, **Barkai N**. Noise-mean relationship in mutated promoters. *Genome Res* 22: 2409–2417, 2012.
37. **Huang S**. Non-genetic heterogeneity of cells in development: more than just noise. *Development* 136: 3853–3862, 2009.
38. **Huber W**, **von Heydebreck A**, **Sultmann H**, **Poustka A**, **Vingron M**. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18: S96–S104, 2002.
39. **Huh I**, **Zeng J**, **Park T**, **Yi S V**. DNA methylation and transcriptional noise. *Epigenetics Chromatin* 6: 9, 2013.
40. **Iyer-Biswas S**, **Jayaprakash C**. Mixed Poisson distributions in exact solutions of stochastic autoregulation models. *Phys Rev E* 90: 052712, 2014.
41. **Johnson TL**, **Ares M**. SMITten by the Speed of Splicing. *Nat Rev Endocrinol* 106: 219–246, 2016.
42. **Jonkers I**, **Lis JT**. Getting up to speed with transcription elongation by RNA polymerase II HHS Public Access. *Nat Rev Mol Cell Biol* 16: 167–177, 2015.
43. **Kanehisa M**, **Sato Y**, **Kawashima M**, **Furumichi M**, **Tanabe M**. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44: D457–D462, 2016.
44. **Karlis D**, **Xekalaki E**. Mixed Poisson Distributions. *Internat Stat Rev* 73: 35–58, 2005.
45. **Kornienko AE**, **Dotter CP**, **Guenzl PM**, **Gisslinger H**, **Gisslinger B**, **Cleary C**, **Kralovics R**, **Pauler FM**, **Barlow DP**. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* 17: 14, 2016.
46. **Kwak H**, **Lis JT**. Control of Transcriptional Elongation. *Annu Rev Genet* 47: 483–508, 2013.
47. **Lappalainen T**, **Sammeth M**, **Friedländer MR**, **'t Hoen PAC**, **Monlong J**, **Rivas MA**, **Gonzàlez-Porta M**, **Kurbatova N**, **Griebel T**, **Ferreira PG**, **Barann M**, **Wieland T**, **Greger L**, **van Iterson M**, **Almlöf J**, **Ribeca P**, **Pulyakhina I**, **Esser D**, **Giger T**, **Tikhonov A**, **Sultan M**, **Bertier G**, **MacArthur DG**, **Lek M**, **Lizano E**, **Buermans HPJ**, **Padioleau I**, **Schwarzmayr T**, **Karlberg O**, **Ongen H**, **Kilpinen H**, **Beltran S**, **Gut M**, **Kahlem K**, **Amstislavskiy V**, **Stegle O**, **Pirinen M**, **Montgomery SB**, **Donnelly P**, **McCarthy MI**, **Flicek P**, **Strom TM**, **Lehrach H**, **Schreiber S**, **Sudbrak R**, **Carracedo Á**, **Antonarakis SE**, **Häsler R**, **Syvänen A-C**, **van Ommen G-J**, **Brazma A**, **Meitinger T**, **Rosenstiel P**, **Guigó R**, **Gut IG**, **Estivill X**, **Dermitzakis ET**, **Dermitzakis ET**. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501: 506–511, 2013.
48. **Lehner B**. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* 4: 170, 2008.
49. **Lehner B**, **Kaneko K**. Fluctuation and response in biology. *Cell Mol Life Sci* 68: 1005–1010, 2011.
50. **Li J**, **Liu Y**, **Kim T**, **Min R**, **Zhang Z**. Gene Expression Variability within and

between Human Populations and Implications toward Disease Susceptibility. *PLoS Comput Biol* 6: e1000910, 2010.
51. **Liu J**, **Wu X**, **Zhang H**, **Pfeifer GP**, **Lu Q**. Dynamics of RNA Polymerase II Pausing and Bivalent Histone H3 Methylation during Neuronal Differentiation in Brain Development. *Cell Rep* 20: 1307–1318, 2017.
52. **Love MI**, **Huber W**, **Anders S**. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550, 2014.
53. **Low TY**, **van Heesch S**, **van den Toorn H**, **Giansanti P**, **Cristobal A**, **Toonen P**, **Schafer S**, **Hübner N**, **van Breukelen B**, **Mohammed S**, **Cuppen E**, **Heck AJR**, **Guryev V**. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep* 5: 1469–78, 2013.
54. **MacNeil LT**, **Walhout AJM**. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* 21: 645–657, 2011.
55. **Martinez-Jimenez CP**, **Eling N**, **Chen H-C**, **Vallejos CA**, **Kolodziejczyk AA**, **Connor F**, **Stojic L**, **Rayner TF**, **Stubbington MJT**, **Teichmann SA**, **de la Roche M**, **Marioni JC**, **Odom DT**. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 355: 1433–1436, 2017.
56. **Mason EA**, **Mar JC**, **Laslett AL**, **Pera MF**, **Quackenbush J**, **Wolvetang E**, **Wells CA**. Gene expression variability as a unifying element of the pluripotency network. *Stem cell reports* 3: 365–77, 2014.
57. **McAdams HH**, **Arkin A**. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* 94: 814–9, 1997.
58. **McCarthy DJ**, **Chen Y**, **Smyth GK**. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 40: 4288–4297, 2012.
59. **Metzger BPH**, **Yuan DC**, **Gruber JD**, **Duveau F**, **Wittkopp PJ**. Selection on noise constrains variation in a eukaryotic promoter. *Nature* 521: 344–7, 2015.
60. **Morgan MD**, **Marioni JC**. CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome Biol* 19: 1–13, 2018.
61. **Müller C**, **Zidek LM**, **Ackermann T**, **de Jong T**, **Liu P**, **Kliche V**, **Zaini MA**, **Kortman G**, **Harkema L**, **Verbeek DS**, **Tuckermann JP**, **von Maltzahn J**, **de Bruin A**, **Guryev V**, **Wang Z-Q**, **Calkhoven CF**. Reduced expression of C/EBPβ-LIP extends health and lifespan in mice. *Elife* 7: e34985, 2018.
62. **Newman JRS**, **Ghaemmaghami S**, **Ihmels J**, **Breslow DK**, **Noble M**, **DeRisi JL**, **Weissman JS**. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441: 840–846, 2006.
63. **Nicolas D**, **Zoller B**, **Suter DM**, **Naef F**. Modulation of transcriptional burst frequency by histone acetylation. *Proc Natl Acad Sci U S A* 115: 7153–7158, 2018.
64. **Nie L**, **Wu G**, **Culley DE**, **Scholten JCM**, **Zhang W**. Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. *Crit Rev Biotechnol* 27: 63–75, 2007.
65. **Nojima T**, **Gomes T**, **Grosso ARF**, **Kimura H**, **Dye MJ**, **Dhir S**, **Carmo-Fonseca M**, **Proudfoot NJ**. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* 161: 526–540, 2015.
66. **Park SJ**, **Song S**, **Yang G-S**, **Kim PM**, **Yoon S**, **Kim J-H**, **Sung J**. The Chemical Fluctuation Theorem governing gene expression. *Nat Commun* 9: 297, 2018.
67. **Parker R**, **Song H**. The enzymes and control of eukaryotic mRNA turnover. *Nat*

Struct Mol Biol 11: 121–127, 2004.
68. **Paulsson J**. Models of stochastic gene expression. *Phys Life Rev* 2: 157–175, 2005.
69. **Pedraza J**, **van Oudenaarden A**. Noise propagation in gene networks. *Science* 307: 1965–1974, 2005.
70. **Potter SS**. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat Rev Nephrol* 14: 479–492, 2018.
71. **Raj A**, **van Oudenaarden A**. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* 135: 216–226, 2008.
72. **Raj A**, **Peskin CS**, **Tranchina D**, **Vargas DY**, **Tyagi S**. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4: e309, 2006.
73. **Raj A**, **Rifkin SA**, **Andersen E**, **van Oudenaarden A**. Variability in gene expression underlies incomplete penetrance. *Nature* 463: 913–918, 2010.
74. **Ran D**, **Daye ZJ**. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Res* 45: e127–e127, 2017.
75. **Rangaraju S**, **Solis GM**, **Thompson RC**, **Gomez-Amaro RL**, **Kurian L**, **Encalada SE**, **Niculescu AB**, **Salomon DR**, **Petrascheck M**. Suppression of transcriptional drift extends C. elegans lifespan by postponing the onset of mortality. *Elife* 4: e08833, 2015.
76. **Raser JM**, **O'Shea EK**. Control of Stochasticity in Eukaryotic Gene Expression. *Science* 304: 1811–1814, 2004.
77. **Ravarani CNJ**, **Chalancon G**, **Breker M**, **de Groot NS**, **Babu MM**. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nat Commun* 7: 10417, 2016.
78. **Rigby RA**, **Stasinopoulos DM**. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C* 54: 507–554, 2005.
79. **Rigby RA**, **Stasinopoulos DM**, **Akantziliotou C**. A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Comput Stat Data Anal* 53: 381–393, 2008.
80. **Risso D**, **Ngai J**, **Speed TP**, **Dudoit S**. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 32: 896–902, 2014.
81. **Sainsbury S**, **Bernecky C**, **Cramer P**. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16: 129–143, 2015.
82. **Sanchez A**, **Choubey S**, **Kondev J**. Regulation of Noise in Gene Expression. *Annu Rev Biophys* 42: 469–491, 2013.
83. **Sanchez A**, **Golding I**. Genetic determinants and cellular constraints in noisy gene expression. *Science* 342: 1188–93, 2013.
84. **Sato K**, **Ito Y**, **Yomo T**, **Kaneko K**. On the relation between fluctuation and response in biological systems. *Proc Natl Acad Sci* 100: 14086–14090, 2003.
85. **Sears KE**, **Maier JA**, **Rivas-Astroza M**, **Poe R**, **Zhong S**, **Kosog K**, **Marcot JD**, **Behringer RR**, **Cretekos CJ**, **Rasweiler JJ**, **Rapti Z**. The Relationship between Gene Network Structure and Expression Variation among Individuals and Species. *PLoS Genet* 11: e1005398, 2015.
86. **Sen P**, **Dang W**, **Donahue G**, **Dai J**, **Dorsey J**, **Cao X**, **Liu W**, **Cao K**, **Perry R**, **Lee JY**, **Wasko BM**, **Carr DT**, **He C**, **Robison B**, **Wagner J**, **Gregory BD**, **Kaeberlein M**, **Kennedy BK**, **Boeke JD**, **Berger SL**. H3K36 methylation promotes longevity by enhancing transcriptional fidelity. *Genes Dev* 29: 1362–76, 2015.
87. **Sharon E**, **van Dijk D**, **Kalma Y**, **Keren L**, **Manor O**, **Yakhini Z**, **Segal E**. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res* 24: 1698–706, 2014.

88. **Sherman MS**, **Lorenz K**, **Lanier MH**, **Cohen BA**. Cell-to-Cell Variability in the Propensity to Transcribe Explains Correlated Fluctuations in Gene Expression. *Cell Syst* 1: 315–325, 2015.
89. **Slieker RC**, **van Iterson M**, **Luijk R**, **Beekman M**, **Zhernakova D V.**, **Moed MH**, **Mei H**, **van Galen M**, **Deelen P**, **Bonder MJ**, **Zhernakova A**, **Uitterlinden AG**, **Tigchelaar EF**, **Stehouwer CDA**, **Schalkwijk CG**, **van der Kallen CJH**, **Hofman A**, **van Heemst D**, **de Geus EJ**, **van Dongen J**, **Deelen J**, **van den Berg LH**, **van Meurs J**, **Jansen R**, **'t Hoen PAC**, **Franke L**, **Wijmenga C**, **Veldink JH**, **Swertz MA**, **van Greevenbroek MMJ**, **van Duijn CM**, **Boomsma DI**, **Slagboom PE**, **Heijmans BT**. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol* 17: 1–13, 2016.
90. **Small EC**, **Xi L**, **Wang J-P**, **Widom J**, **Licht JD**. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proc Natl Acad Sci U S A* 111: E2462-71, 2014.
91. **Stasinopoulos MD**, **Rigby RA**, **Heller GZ**, **Voudouris V**, **Bastiani F De**, **Rigby RA**, **Heller GZ**, **Voudouris V**, **Bastiani F De**. Flexible Regression and Smoothing. CRC Press, 2017.
92. **Sultan M**, **Amstislavskiy V**, **Risch T**, **Schuette M**, **Dökel S**, **Ralser M**, **Balzereit D**, **Lehrach H**, **Yaspo M-L**. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* 15: 675, 2014.
93. **Suter DM**, **Molina N**, **Gatfield D**, **Schneider K**, **Schibler U**, **Naef F**. Mammalian genes are transcribed with widely different bursting kinetics. *Science* 332: 472–4, 2011.
94. **Swain PS**, **Elowitz MB**, **Siggia ED**. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci* 99: 12795–12800, 2002.
95. **Tan Q**, **Heijmans BT**, **Hjelmborg JVB**, **Soerensen M**, **Christensen K**, **Christiansen L**. Handling blood cell composition in epigenetic studies on ageing. *Int J Epidemiol* 46: 1717–1718, 2017.
96. **Thattai M**. Universal Poisson Statistics of mRNAs with Complex Decay Pathways. *Biophys J* 110: 301–305, 2016.
97. **Thattai M**, **van Oudenaarden A**. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci* 98: 8614–8619, 2001.
98. **Tirosh I**, **Barkai N**. Two strategies for gene regulation by promoter nucleosomes. *Genome Res* 18: 1084–91, 2008.
99. **de Torrente L**, **Zimmerman S**, **Taylor D**, **Hasegawa Y**, **Wells CA**, **Mar JC**. *pathVar:* a new method for pathway-based interpretation of gene expression variability. *PeerJ* 5: e3334, 2017.
100. **Vaquero-Garcia J**, **Barrera A**, **Gazzara MR**, **González-Vallinas J**, **Lahens NF**, **Hogenesch JB**, **Lynch KW**, **Barash Y**. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 5: e11752, 2016.
101. **Viñuela A**, **Brown AA**, **Buil A**, **Tsai P-C**, **Davies MN**, **Bell JT**, **Dermitzakis ET**, **Spector TD**, **Small KS**. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum Mol Genet* 27: 732–741, 2018.
102. **Voisin S**, **Eynon N**, **Yan X**, **Bishop DJ**. Exercise training and DNA methylation in humans. *Acta Physiol (Oxf)* 213: 39–59, 2015.
103. **Wan Y**, **Larson DR**. Splicing heterogeneity: separating signal from noise. *Genome Biol* 19: 86, 2018.
104. **Wang Z**, **Zhang J**. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc Natl Acad Sci* 108: E67–E76, 2011.

105. **White RR**, **Milholland B**, **MacRae SL**, **Lin M**, **Zheng D**, **Vijg J**. Comprehensive transcriptional landscape of aging mouse liver. *BMC Genomics* 16: 899, 2015.
106. **Wolf L**, **Silander OK**, **van Nimwegen E**. Expression noise facilitates the evolution of gene regulation. *Elife* 4: e05856, 2015.
107. **Wu L**, **Fan J**, **Belasco JG**. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci* 103: 4034–4039, 2006.
108. **Wu S**, **Li K**, **Li Y**, **Zhao T**, **Li T**, **Yang Y-F**, **Qian W**. Independent regulation of gene expression level and noise by histone modifications. *PLOS Comput Biol* 13: e1005585, 2017.
109. **Yamashita A**, **Chang T-C**, **Yamashita Y**, **Zhu W**, **Zhong Z**, **Chen C-YA**, **Shyu A-B**. Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nat Struct Mol Biol* 12: 1054–63, 2005.
110. **Yip SH**, **Sham PC**, **Wang J**. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief. Bioinform.* bby011, 2018.
111. **Yu Y**, **Fuscoe JC**, **Zhao C**, **Guo C**, **Jia M**, **Qing T**, **Bannon DI**, **Lancashire L**, **Bao W**, **Du T**, **Luo H**, **Su Z**, **Jones WD**, **Moland CL**, **Branham WS**, **Qian F**, **Ning B**, **Li Y**, **Hong H**, **Guo L**, **Mei N**, **Shi T**, **Wang KY**, **Wolfinger RD**, **Nikolsky Y**, **Walker SJ**, **Duerksen-Hughes P**, **Mason CE**, **Tong W**, **Thierry-Mieg J**, **Thierry-Mieg D**, **Shi L**, **Wang C**. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat Commun* 5: 3230, 2014.
112. **Zaugg JB**, **Luscombe NM**. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. *Genome Res* 22: 84–94, 2012.
113. **Zhang H-M**, **Diaz V**, **Walsh ME**, **Zhang Y**. Moderate lifelong overexpression of tuberous sclerosis complex 1 (TSC1) improves health and survival in mice. *Sci Rep* 7: 834, 2017.
114. **Zhang Z**, **Qian W**, **Zhang J**. Positive selection for elevated gene expression noise in yeast. *Mol Syst Biol* 5: 299, 2009.
115. **Zhao S**, **Fung-Leung W-P**, **Bittner A**, **Ngo K**, **Liu X**. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 9: e78644, 2014.
116. **Zoller B**, **Nicolas D**, **Molina N**, **Naef F**. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol Syst Biol* 11: 823, 2015.