Article type: Review

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

Authors: Tristan V. de Jong¹, Yuri M. Moshkin²,³, Victor Guryev¹

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.
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[RNA] = Var[RNA] = μ) represented by a constant ratio of synthesis to degradation rates: μ = λ/γ = 〈λ〉. Gene expression noise, expressed through a squared coefficient of variation in RNA copy number, is reciprocal to the mean of RNA copy number: \[ cv^2(RNA) = \frac{Var[RNA]}{E[RNA]^2} = \mu^{-1} \] (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 (p_on) is a Beta-distributed random variable, which depends on RNA degradation rate normalized \[ \hat{k}_{on} = \frac{k_{on}}{\gamma} \] and \[ \hat{k}_{off} = \frac{k_{off}}{\gamma} \] rates of promoter switching: \[ p_{on} \sim Beta(\hat{k}_{on}, \hat{k}_{off}) \]. This, in turn, defines the
distribution of otherwise constant Poisson rate \( \mu = \lambda p_{\text{on}} \) 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 \( E[\text{RNA}] = \lambda E[p_{\text{on}}] = \langle \mu \rangle \) and \( Var[\text{RNA}] = \langle \mu \rangle + Var[\mu] = \langle \mu \rangle + \langle \mu \rangle^2 Var[p_{\text{on}}] \), from where \( cv^2(\text{RNA}) = \langle \mu \rangle^{-1} + cv^2(\mu) = \langle \mu \rangle^{-1} + cv^2(p_{\text{on}}) \) \( \text{Fig. 1} \). 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 \( (cv^2(\mu) = cv^2(p_{\text{on}})) \).

Under limiting conditions of \( \tilde{k}_{\text{off}} \gg k_{\text{on}} \) and \( \tilde{k}_{\text{off}} \gg 1 \), i.e. when a gene is transcribed in short bursts, the \( p_{\text{on}} \) distribution converges to Gamma \((p_{\text{on}} \sim \text{Gamma}(k_{\text{on}}, \tilde{k}_{\text{off}}))\) 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 \( \langle \mu \rangle^{-1} = (bf_{\text{b}})^{-1} \) and non-Poisson noise is \( cv^2(\mu) = cv^2(p_{\text{on}}) = f_{\text{b}}^{-1} \), where \( b = \lambda k_{\text{off}}^{-1} \) is a burst size and \( f_{\text{b}} = \tilde{k}_{\text{on}} \) 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
with promoter fluctuation, RNA synthesis or degradation kinetics is expected to modulate the within-cell and cell-to-cell variability in RNA copy number and thus gene noise.

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

The amount of RNA observed in a cell is the consequence of equilibrium between synthesis and degradation. This means not only fluctuations in the synthesis rate, but also variations in the degradation rate are likely to influence both the average expression as well as the variation in expression (57, 97). The half-life of RNA molecules depends on the length of the 3’-poly(A)-tail which is removed through deadenylation prior to degradation (67, 109). As a direct consequence of the two-state promoter model, the total gene expression noise (Poisson and non-Poisson) is directly proportional to the RNA degradation rate. This implies an increased noise level for RNA species with shorter half-life and a decreased noise for the stable RNA molecules. For example, the presence of certain microRNAs have been shown to increase the rate of RNA deadenylation (107) and one can predict that such a mechanism will turn up the gene noise. Strikingly, although RNA synthesis and degradation, at first glance, are two independent processes, the RNA degradation rate was found to be regulated by transcription (13, 33). In terms of gene noise, the existence of a coupling between synthesis and degradation rates has a profound consequence as it leads to non-Poisson RNA “birth-death” process even in the absence of upstream cellular drives (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 \( \text{CRCX4} \) 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 \( \text{Pou5f1} \) (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 \( \text{C.elegans} \) 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 ($\sigma^2$), the (squared) coefficient of variation ($cv^2$, 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 \( \mu_X \) and \( \sigma_X^2 \) are original mean and variance respectively. 

Among commonly used VSTs are the logarithm (\( \log(X) \)) and generalized logarithm (\( g\log_2(X) = \log(X + \sqrt{X^2 + 1}) \)) functions (38). This implies that the variance of \( \log_2 \) or \( g\log_2 \) transformed variables corresponds to the squared coefficient of variation of the original variable (\( cv_X^2 \)) as 

\[
Var[\log_2(X)] \approx \log(2)^{-2} \frac{\sigma_X^2}{\mu_X^2} = \log(2)^{-2} cv_X^2
\]

\[
Var[g\log_2(X)] \approx \log(2)^{-2} \frac{\sigma_X^2}{\mu_X^2 + 1} \approx \log(2)^{-2} cv_X^2 \text{ (for } \mu_X \gg 1 \text{). Thus, it makes no sense to estimate neither } cv \text{ nor Fano factor for VST transformed variables as their variance is already proportional to } cv_X^2. \text{ Similar logic applies to robust measures of variability as } MAD[\log_2(X)] \approx \text{ median}(|\log_2(X_i/\bar{X})|) \text{ and } MAD[g\log_2(X)] \approx \text{ median}(|g\log_2(X_i/\bar{X})|) \text{ (for } X_i \gg 1 \text{), and additional normalization of } MAD \text{ 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 \( \mu_X \) and \( \alpha_X \) 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: $\text{Var}(X) = \mu \rho$, 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 ($\mu$), 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 ($\mu$). 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: $\text{Var}[X] = \mu + \alpha \mu^2$, $\text{cv}^2(X) = \mu^{-1} + \alpha$ respectively (44, 79). Non-Poisson variation $- \alpha$ 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 \( -\mu\), while that of differential gene noise analysis is to estimate overdispersion \( -\alpha\) 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 \( (\mu^{-1})\) and non-Poisson \( (\alpha)\) 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 \( (\mu)\) and non-Poisson variation \( (\alpha)\) 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: \( X \sim ind NB(\mu, \alpha) \) is defined as:

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

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:
\[ \log(X_i) \sim \text{age}_j \beta_{\mu_j} + \log(N_i), \]
\[ \log(\alpha) \sim \text{age}_j \beta_{\alpha_j}, \]
where \( i = 1, \ldots, n \) is \( i \)th observation of gene's mRNA counts \((X_i)\); \( j = 1, \ldots, p \) 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 \( \text{age}_j \), on library size \((N_i)\) normalized mean mRNA counts \((\mu_j = e^{\beta_{\mu_j}}, \text{cpm}_j = 10^6 \mu_j)\). 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 \((\alpha)\), where \( \beta_{\alpha_j} \) is a maximum-likelihood estimation of overdispersion parameter \((\alpha_j = e^{\beta_{\alpha_j}})\).

Significance values of age-mediated changes in \( \mu \) and \( \alpha \) 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:

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

Where the reduced model omits factor effect (age) from the model of \( \mu \): \( \log(X_i) \sim \beta_{\mu_0} + \log(N_i) \), 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:

\[ \log(X_i) \sim \text{age}_j \beta_{\mu_j} + \log(N_i), \]
\[ \log(\alpha) \sim \beta_{\alpha_0} \]

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

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

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

\[ D_{\alpha} = -2 \log \frac{\text{likelihood for GLM model}}{\text{likelihood for GAMLSS model}} = -2 \log \frac{\mathcal{L}(\mu_j, \alpha_0 | X_i)}{\mathcal{L}(\mu_j, \alpha_j | X_i)}. \]
\( D_\mu, D_{\mu_{\text{GLM}}} \) and \( D_\alpha \) are asymptotically \( \chi^2 \)-distributed with degrees of freedom equal to a difference between the number of compared models’ parameters. Thus, from this example it is clear that GAMLSS is an extension of a GLM model allowing for the estimation of factor effects on both parameters of the distribution of RNA counts, namely mean and overdispersion (non-Poisson noise).

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

Having estimated the parameters \( \mu \) and \( \alpha \) for liver genes expressed in young and old mice, we noted that their absolute values were practically uncorrelated \((\rho(\mu, \alpha) \rightarrow 0)\). This could be attributed directly to the given parametrization of the \( NB(\mu, \alpha) \), which implies an asymptotic independence of the estimated parameters. It follows from the Fisher information matrix as its element \( I_{\mu \alpha} = -E_{\mu, \alpha} \frac{\partial^2}{\partial \mu \partial \alpha} \log(P(X| \mu, \alpha)) = 0 \). To that, changes in the mean gene expression and the non-Poisson variation occurring with age were also almost uncorrelated \((\rho(\Delta \mu, \Delta \alpha) \rightarrow 0)\). Testing under the assumption that the cellular RNA concentration (total number of RNA molecules per cell) is the same for the samples taken from young and old mice, we scored about a comparable number of genes for which the mean RNA counts either increased or decreased significantly with age (Fig. 2A, 3A).

Estimation of the mean also yielded the estimation of the Poisson variation as they are reciprocal to each other (Poisson variation = \( \mu^{-1} \)). In contrast to the Poisson variation, non-Poisson variation increased with age (Fig. 2B). Importantly, applying the GAMLSS model enabled for the identification of genes for which the non-Poisson variation, but not the mean, changed significantly with age (Fig. 2B, 3B).

However, it must be noted that the relative standard errors of overdispersion estimates 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 $\mu_0$ (counts) and $\alpha_0$ (non-Poisson noise) for reference samples and fold changes ($\delta$) 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 ($\mu_0$: $\{10, 100, 1000\}$ and $\alpha_0$: $\{0.1, 0.25, 0.5\}$) 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 $\mu_0$ 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
expectation ($\mu$) of some measurable quantity ($X$) in response to an influence is proportional to its initial variance: $|\mu_1 - \mu_0| \sim \text{Var}(X)$. However, this relationship holds only true for Gaussian-like distributed quantities under the assumption of a fixed variance: $\text{Var}(X_1) \sim \text{Var}(X_0)$. Nonetheless, if log transformed RNA counts approximate a Gaussian-like distribution, then the fluctuation-response relationship takes on the following form: $|\log(\mu_1/\mu_0)| \sim \alpha = bcv^2$, as a result of the Taylor expansions for the moments for genes expressed at large copy number ($\mu \gg 1$). We noted a modest, but significant, positive correlation between absolute log 2 fold changes in the mean gene expression for old and young mice with non-Poisson variation for young mice (Fig. 5A). A lack of a stronger correlation could be due to the violation of the fluctuation-response assumption of a fixed variance or overdispersion for log-transformed variables. In general, this substantiates the fluctuation-response relationship for the RNA copy number.

Estimates of gene variation from tissues retain information on gene state determinants of non-Poisson noise.

Finally, we wondered if the estimate of non-Poisson variation from RNA-sequencing data of cell populations contain information on gene state determinants. To this end, we compared the genes’ non-Poisson variation estimates with their promoter DNA-sequence composition. First, we noted that on average, that the non-Poisson variation was higher for genes that were regulated by TATA-containing promoters (Fig. 5B). Second, in accordance with the fluctuation-response relationship (Fig. 5A), aging induced more pronounced changes in the mean expression of genes with TATA-containing promoters (Fig. 5B, C). Overall, this result is in agreement with the TATA-mediated promoter fluctuation caused by a competition between distinct TBP-co-activator complexes (77, 82, 87) and it substantiates that gene state signals are retained in cell population estimates of non-Poisson variation.

To conclude this brief showcase of GAMLSS, we advocate for the use of this framework to dissect the determinants of both the mean RNA counts and the non-Poisson variation as two independent parameters of gene expression network.

Combining other -omics data with RNA-seq can lead to new discoveries.

A connection between the gene expression variability measured on different levels: cell-to-cell, inter-individual and inter-population has been suggested previously (23, 25). The rapid development of accessible and cost-efficient methods for single-cell RNA-seq (scRNA-
seq) will provide us with improved estimates of cell-to-cell variability in gene expression (70). Flow cytometry techniques can help in the further separation into (so called / the suggested) macro-heterogeneity, which is the variability that encompasses both the on- and off-state of genes, as well as the micro-heterogeneity, which represents the variability in gene expression of genes in different (37). Further, recently generated large transcriptome datasets for hundreds of individuals (2, 47) should increase our understanding of transcriptome variability at population level.

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

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

To conclude, the analysis of differential transcriptome variability complements the standard analysis of differential gene expression and reveals another dimension of expression analysis. With the further development of tools and with a wider acceptance of these methods, we will advance our understanding of the mechanisms underlying the regulation of transcription, common physiological traits and disease predispositions.
Table 1. Commonly used measures of variability

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient of variation (signal to noise ratio)</td>
<td>$cv = \sigma / \mu$</td>
</tr>
<tr>
<td>Fano factor (noise strength)</td>
<td>$F = \sigma^2 / \mu$</td>
</tr>
<tr>
<td>Median Absolute Deviation from the Median</td>
<td>$MAD = \text{median}(</td>
</tr>
<tr>
<td>Coefficient of Dispersion</td>
<td>$COD = MAD / \bar{X}$</td>
</tr>
<tr>
<td>Quartile Coefficient of Dispersion</td>
<td>$QCOD = (Q_3 - Q_1) / (Q_3 + Q_1)$</td>
</tr>
</tbody>
</table>

$\bar{X}$ - 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 $\mu$ is a random variable distributed with expectation $\langle \mu \rangle$ and variance $\text{Var} [\mu]$. Key equations for the expectation ($E[RNA]$), variance ($\text{Var}[RNA]$) 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 log2 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 panel...
Genes for which the non-Poisson noise increased or decreased significantly with age are marked in red or blue respectively.

C) Heatmap depicting a power analysis of the likelihood ratio (LR) test for fold changes ($\delta$) in $\mu_0$ (mean counts). For each power analysis (1000) pairs of samples from reference $NB(\mu_0, \alpha_0)$ and test $NB(\delta \mu_0, \alpha_0)$ distributions were simulated with $\mu_0 \in \{10,100,1000\}$, $\alpha_0 \in \{0.1,0.25,0.5\}$ and $\delta \in \{\frac{1}{4}, \frac{1}{3}, ..., 3, 4\}$. Sample sizes were $\{5,10,...,100,1000\}$. Null hypothesis: $H_0: \mu_0 = \delta \mu_0$ were rejected at significance level of 0.05 and power was calculated as the probability of rejecting $H_0$. Red indicates high power, white -low.

D) Heatmap depicting a power analysis of the likelihood ratio (LR) test for fold change ($\delta$) in $\alpha_0$ (non-Poisson noise).

Figure 3. Examples of differentially expressed genes (A) and genes showing increase in non-Poisson variability with age (B). Upper panel, boxplots of selected liver genes’ mRNAs copy numbers (expressed as log$_2$(cpm)) for young (green, $n = 6$) and old (red, $n = 6$) C57BL/6J mice. Whiskers extend to minimum and maximum values. Middle panel, boxplots of log$_2$(cpm) residual values corrected for genes’ grand mean expression for young and old mice (~ gene). Lower panel, boxplots of log$_2$(cpm) residuals corrected for genes’ group-wise mean expression in young and old mice (~gene:age). The middle panel serves to illustrate differential gene expression, while the lower panel shows whether the gene expression variability is affected by age. Genes were selected based on significance of the age-mediated changes in mean mRNA counts (A, FDR$_{cpm} \leq 0.05$) or changes in non-Poisson variability (B, FDR$_{non-Pois. variability} \leq 0.05$). For (B), note an increase in log$_2$(cpm) variability for selected genes in population of 20 weeks old mice due to an increase in non-Poisson variability with age as compared to 5 weeks mice. Left panel in (B) shows genes associated with complement and coagulation cascades according to KEGG annotation, the right panel shows a selection of 30 genes with the highest statistically significant gain in non-Poisson variability.

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 log₂ 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


50. Li J, Liu Y, Kim T, Min R, Zhang Z. Gene Expression Variability within and


Sanchez Sainsbury by RNA polymerase II. Analysis of control genes or samples.


1000

1001

1002

1003

1004

1005

1006


System state variables:
- cell cycle / growth
- metabolism
- aging
- etc.

Regulatory variables:
- activator / repressor binding
- nucleosome assembly
- histone code
- DNA methylation

Gene state variables:
- Upstream / non-Poisson:
  - promoter fluctuation: $k_{on} / k_{off}$
- Downstream / Poisson:
  - RNA synthesis rate: $\lambda$
  - RNA degradation rate: $\gamma$

Gene expression parameters:
- Expectation: $E[RNA] = (\mu)$
- Variance: $\text{Var}[RNA] = (\mu) + (\mu)^2 b^{-1}$
- Noise: $\text{cv}^2[RNA] = (\mu)^2 + (\mu)^2 b^{-1}$

Two-state promoter model:
- burst size & freq.: $b = \lambda / k_{on} ; f_b = k_{on} / \gamma$
- $E[RNA] = (\mu) = b f_b$
- $\text{Var}[RNA] = (\mu) + (\mu)^2 b^{-1}$
- $\text{cv}^2[RNA] = (\mu)^2 + (\mu)^2 b^{-1}$

- expectation: $E[RNA] = (\mu)$
- variance: $\text{Var}[RNA] = (\mu) + (\mu)^2 b^{-1}$
- noise: $\text{cv}^2[RNA] = (\mu)^2 + (\mu)^2 b^{-1}$

Downloaded from www.physiology.org/journal/physiolgenomics at Biblio der Rijksuniversiteit (129.125.166.130) on April 8, 2019.
A

B

C

D

NB(\mu_0, \alpha_0) vs. NB(\delta \mu_0, \alpha_0):

\text{counts}

\text{fold change}

\text{power} = P(\text{reject } H_0 | \text{true } H_1)

\text{non-Poisson noise: log}_2(\text{old / young})

\text{cpm: log}_2(\text{old / young})

\text{FDR < 0.05}
**A**

- Scatter plot showing non-Poisson noise: \( \log_{2}(\text{old/young}) \)

- Correlation coefficient: \( r = 0.299 \)

**B**

- Bar chart showing KEGG model coefficients for various biological pathways:
  - Complement and coagulation cascades
  - Val, Leu, Ile degradation
  - Chemokine signaling pathway
  - Adrenergic signaling in cardiomyocytes
  - Platelet activation
  - Carbon metabolism
  - Cytokine-cytokine receptor interaction
  - Natural killer cell mediated cytotoxicity
  - Peroxisome
  - Drug metabolism by other enzymes
  - Necrosis
  - Pertussis
  - Nucleotide excision repair
  - Acute myeloid leukemia
  - Protein processing in ER
  - Metabolism by cytochrome P450
  - Trp metabolism
  - Parkinson disease
  - Legionellosis
  - DNA replication

KEGG model coef. values range from 0 to 0.3.
A

\[ r = 0.374 \]

B

TATA - less

TATA - containing

C

\[ r = 0.92, p < 0.001 \]