INTERINDIVIDUAL GENE EXPRESSION VARIABILITY RESULTS FROM DNA SEQUENCE ENCODED GENE NOISE DETERMINANTS ADJUSTED TO FLUCTUATION IN SYSTEM-STATE

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

Manuscript in preparation
ABSTRACT

Cell-to-cell variability in gene expression originates from a doubly-stochastic Poisson process determined by fluctuations in Poisson “birth/death” rate. However, from the law of large numbers information on the kinetic gene noise determinants is expected to vanish from cell population RNA counts. Here we counter this argument both on theoretical and empirical grounds. First, we showed that as a result of fluctuation-response relationship theorem distribution of RNA counts sampled from cell populations contains such information. Second, overdispersion in bulk RNA sequencing counts can be predicted with unexpectedly high accuracy from DNA sequence context. As such, we noted that overdispersion increases for genes with higher GC content downstream of promoter. We explain this by a non-productive accumulation of RNA polymerase II on such genes due to higher melting energy costs. Third, certain DNA and chromatin signatures are common for both single-cell and cell population gene noise. This includes over-representation of TATA-box and promoter invasion by nucleosome for genes with increased overdispersion in RNA counts. Finally, we showed that metabolic and aging cues modulate fluctuations in both system-state and gene-state parameters, thus tuning interindividual gene expression variability. From these, we conclude that interindividual variability in RNA copy number stems from DNA encoded gene noise and environmental fluctuations.
INTRODUCTION

Due to the stochastic nature of RNA synthesis and degradation, the expression of genes is inherently noisy. The observed variation in gene expression originates from probabilistic fluctuations in the system-state (gene-network) and cell-state (environmental) variables. The former is usually referred to as “intrinsic” gene noise, while the latter as “extrinsic” gene noise (Elowitz, Levine, Siggia, & Swain, 2002; Hilfinger & Paulsson, 2011; Raser & O’Shea, 2004; Swain, Elowitz, & Siggia, 2002; Thattai & van Oudenaarden, 2001).

“Intrinsic” gene noise depends on the statistics of promoter switching, RNA synthesis, processing- and degradation rates. Consequently, the architecture of a promoter region, the epigenetic state of a gene (histone-code, DNA methylation) and the mRNA sequence/structure-dependent lifetime all modulate the “intrinsic” gene noise (Carey, van Dijk, Sloot, Kaandorp, & Segal, 2013; Faure, Schmiedel, & Lehner, 2017; A. Sanchez & Golding, 2013; A Sanchez, Garcia, Jones, Phillips, & Kondev, 2011; Sharon et al., 2014; Wu et al., 2017). The presence of a strong TATA box, the multiplicity of transcription factor binding sites (TFBS) as well as the assembly/remodelling of nucleosomes on the promoter region all increase expression noise (Blake et al., 2006; Hornung et al., 2012; Raser & O’Shea, 2004; Alvaro Sanchez, Choubey, & Kondev, 2013; Sharon et al., 2014; Suter et al., 2011; Tirosh & Barkai, 2008; To & Maheshri, 2010). Mechanistically, a strong TATA box attracts distinct TBP-co-activator complexes which increases the stochasticity of the promoter pulsing. A weak TATA box favors the stable binding of TBP-TFIID, which in turn lowers the promoter stochasticity (Ravarani, Chalancon, Breker, de Groot, & Babu, 2016). The multiplicity of binding sites and nucleosome assembly within the promoter region can both disturb the TF binding kinetics causing increased promoter activity fluctuations causing an increase in noise (Lin & Buchler, 2018; A. Sanchez & Golding, 2013; Sharon et al., 2014). These emphasize the partially deterministic nature of “intrinsic” gene noise, which is defined by the promoter DNA sequence. “Extrinsic” gene noise is driven by a plethora of intra- and extracellular factors. These include cell-to-cell and cell-state dependent variations in the
concentration of both RNA polymerase and transcription factors, as well as cell cycle alterations in gene copy number and fluctuations in the cell’s metabolic/energetic state (Bahar et al., 2006; Hensel et al., 2012; Kiviet et al., 2014; Shahrezaei & Marguerat, 2015; Shahrezaei, Ollivier, & Swain, 2008; Volfson et al., 2006). However, due to the coupling between system- and cell-state variables it is difficult to distinguish between the “intrinsic” and “extrinsic” gene noise (Paulsson 2005; Hilfinger and Paulsson 2011; Sherman et al. 2015).

The “birth-death” of RNA molecules universally follows stochastic Poisson process under assumptions that promoter is constantly active and that RNA degradation rate is independent of synthesis rate (Thattai 2016). Naturally, however, transcription occurs in bursts with system-state variables being dependent on cell-state (Suter et al. 2011; Dar et al. 2012; Zoller et al. 2015). This leads to double-stochastic Poisson process, a.k.a. a mixed Poisson process, where the gene “birth/death” rate ($\mu$) is stochastic itself (Iyer-Biswas and Jayaprakash 2014; Dattani and Barahona 2017; Park et al. 2018). For a mixed Poisson processes, the noise is represented by the squared coefficient of variation ($cv^2$) which partitions into the Poisson and non-Poisson parts. The former is given by the inverse expectation of the “birth/death” rate ($E(\mu)^{-1}$), while the latter equals the $cv^2(\mu)$ (Supplementary note 1). In other words, the non-Poisson noise reflects fluctuations in the Poisson rate caused by cell population heterogeneity and upstream cellular drives, such as transcriptional bursts (Dattani and Barahona 2017). Thus, we define the “extrinsic” gene noise as a non-Poisson component of the total gene noise, which can be predicted from a fixed system-state parameter such as the promoter/gene DNA sequence context.

Here we assessed the molecular and biological determinants of the non-Poisson gene expression noise in laboratory mice and rats’ tissues (Munger et al., 2014; Yu et al., 2014). Applying the Generalized Additive Model for Location, Scale and Shape (GAMLSS) framework (Stasinopulos et al., 2017) we estimated non-Poisson gene expression noise from mRNA counts of mouse Diversity Outbred (DO) strain genes. For the mouse genes estimates of non-Poisson noise were correlated to genetic and
epigenetic factors from publicly available database. Estimates of the biological coefficient of variation (BCV\(^2\)) were also correlated between the mouse genes expressed in the liver and rat genes expressed in various tissues. The resulting correlations suggest the existence of an inherent cell-state independent driver for non-Poisson gene expression noise. We showed that a genes’ relative BCV\(^2\) could be predicted with a high accuracy from the DNA sequence context based on regions flanking the transcription start site (TSS) for both mouse and rat genes. For inherently noisy genes, we noted a marked increase in DNA duplex stability downstream of TSS due to elevated GC content. This leads to non-productive accumulation/stalling of RNA polymerase II (RNAP) at the beginning of a gene. Finally, the resultant magnitude of non-Poisson noise is modulated genome-wide by extrinsic biological factors, such as ageing or the dietary regime. Thus, although Poisson and non-Poisson components of gene noise are conditioned on extrinsic factors (Sherman, Lorenz, Lanier, & Cohen, 2015), we showed that non-Poisson gene noise is, to a large degree, determined by the DNA sequence context.

**RESULTS**

The total RNA-content within a cell at any time is dependent on a stochastic Poisson process which depend on synthesis and degradation rates (supplemental equations 1 and 2). Upstream cellular drives, like promotor switching, transcription factor binding, and other micro-environmental fluctuations increase stochasticity of this process, resulting in a double-stochastic Poisson process (s.eq. 3). These variations in expression cause the cell-to-cell RNA counts to follow a mixed-Poisson distribution (s.eq. 4).

Due to this formulation, the total noise (variance) of counts can be separated into two parts, the Poisson and non-Poisson noise/variability (s.eq. 5). Poisson noise is referred to as “intrinsic”, as it is dependent on the synthesis and degradation part of the equation. The non-Poisson noise is defined as “extrinsic”, as it represents cell-to-cell
variability in RNA-copy number. However, this distinction is ambiguous as both are influenced by upstream cellular drives (Paulsson, 2005).

To understand how the Poisson part of the equation befits the synthesis rate and the non-Poisson part befits the upstream cellular drives we can equate it to the activation rate and inactivation rate as a simple two state promotor model:

$$\text{OFF} \xrightleftharpoons[k_{\text{off}}]{k_{\text{on}}} \text{ON} \rightarrow \lambda \text{RNA} \rightarrow \emptyset,$$

In which $k_{\text{on}}$ represents the rate of the promotor switching to the on state, and thus transcription taking place, and $k_{\text{off}}$, the rate of the promoter switching to the off state. As upstream cellular drives influence $k_{\text{on}}$ and $k_{\text{off}}$, which in turn influence the RNA production through the overall synthesis rate (s.eq. 6). These dynamics translate to the burst size (time of $k_{\text{on}} \times$ synthesis rate) and the burst frequency (rate of switching to $k_{\text{on}}$). The distributions of activity of these hypothetical promotors follow a beta distribution. The linear act of synthesis and degradation are expected to result in a distribution of RNA molecules which adheres to a Poisson distribution. The switching between the ON and OFF states, initiating transcription adds another layer of variation, resulting in the mixed Poisson distribution. Under conditions in which a promoter would alternate fast between the ON and OFF states, or if the promoter is constantly ON, then the RNA counts should hypothetically return to a Poisson distribution. (s.eq. 6 and 7). This means the synthesis of RNA is a double stochastic process for which the Poisson rate and the characteristic gene noise can be separated.

**Single cell vs inter-individual variability**

In principle, it should not be possible to detect cell state fluctuations ($\eta \zeta$) in the non-Poisson noise derived from bulk RNA-seq experiments. This is because the observed amount of RNA counts stems from the sum of molecules taken from many cells, this will mask the non-Poisson variation between cells (s.eq. 8). However, biological systems are open and are affected by system state fluctuations, which act similarly on all cells within a cell population (tissue). Due to the interference of these fluctuations
the effects are observed among bulk samples and here we prove these are related to the cell state fluctuations ($\eta \zeta$) through a fluctuation response relationship. We state that the biological coefficient of variation ($BCV^2$) observed between individuals is dependent on both fluctuations in the system state and characteristic gene noise.

The fluctuation response relationship dictates that the response to a change of system state variable ($\Delta a$) is proportional to its initial variance (s.eq. 10). This means that if a change in a system state alters the noise of a given gene, it will always be proportional to its initial noise, in this case, the characteristic gene noise. Due to this relationship, variation caused by upstream cellular drives would be carried to variation between tissues due to variation in the system state among samples.

The fluctuation response relationship requires that three assumptions are met: I) changes in $x$ (RNA counts of a given gene in our case), must be linearly coupled to the change in the system state ($\Delta a$); II) the relationship must be Gaussian like, and III) the change in system state ($\Delta a$), must be small. Cell-to-cell RNA counts ($x$) are generated by a double-stochastic Poisson process as discussed before, which means that their log transformed distributions will be Gaussian like (s.eq. 11). Applying the log transformation to the RNA-seq distribution allows us to model the distribution of RNA counts as a fluctuation response relationship for single cell samples (s.eq. 12). This means that for a given individual the expected RNA-counts are distributed around a single average with a distribution based on the system state variable and the characteristic gene noise (s.eq. 13). It is expected that between individuals, the system state variable ($a$), is distributed around a mean, and the characteristic gene noise ($\eta \zeta$) varies between individuals independently of the system state variables. These distributions can be treated as two random variables from which the moments can be calculated, resulting in an expected and variance (s.eq. 14). The resulting observed variation between individuals is then a linked fluctuation variable containing information on both characteristic gene noise and system state variation between individuals. Calculating the non-Poisson noise or $BCV^2$ between individuals thus
corresponds to this linked variable, which in turn represents noise caused by system state variables driven by cell-to-cell noise caused by the characteristic gene noise (s.eq. 15).

**Diversity outbred mice as model for studying expression overdispersion**

The question which stems from the fluctuation response equation is which intrinsic factors- or upstream cellular drives can be measured from bulk RNA-seq data. To answer this question we retrieved 1086 bulk RNA-sequencing libraries taken from mouse liver samples from diversity outbred (DO) and inbred mouse taken at 20 or 24 weeks from the Gene Expression Omnibus (Gu et al., 2016). The DO mice were used as an example of a heterogeneous population with free access to either standard rodent chow containing 6% fat by weight (73 females and 68 males) or high-fat chow containing 22% fat by weight (70 females and 66 males).

Because for cell population it is merely impossible to derive the exact analytical form of a mixing distribution, we compared goodness-of-fit (GOF) of several mixed Poisson distributions to model mRNA copy number of genes’ expressed in the liver of Diversity Outbred (DO) mouse strain at 26 weeks of age (Munger et al., 2014). Expectedly, mRNA counts were better fitted by mixed Poisson distributions than by a Poisson distribution. Three parameter Sichel and Delaporte distributions had no advantage over two parameter NB and Inverse Gaussian-Poisson (IGP) distributions (Rigby et al. 2008) (Figure S1A). As GOFs were comparable for NB and IGP models of mRNA counts and, because, NB could be related to the two-state promoter model (Dattani & Barahona, 2017; Raj, Peskin, Tranchina, Vargas, & Tyagi, 2006), we applied NB model to estimate non-Poisson noise (BCV²) of mRNA copy number.

Estimation of non-Poisson noise (BCV²) in mRNA counts could be affected by technical variation in RNA-seq experiments. However, for most of the genes overdispersion caused by technical replication was negligibly small (Figure S1A). Thus, for simplicity, to model mRNA counts we did not correct for technical replicates as a random-effect variable. To that, we excluded from the analysis genes with “salt and
pepper” expression in biological replicates as excess of zeroes or low mRNA counts might bias the estimation of overdispersion. As a result, the inclusion of lowly expressed genes causes a negative correlation between otherwise independent parameters of NB distribution: mean and overdispersion (Figure S1B). The use of left-side truncated NB distribution lowered a correlation between mean and overdispersion estimates (Figure S1C), while for genes expressed constitutively across biological replicates these two parameters were uncorrelated (Figure S1D).

**DNA determinants of expression variability**

The most basal and intrinsic factor when thinking of transcriptional mechanisms is the genome sequence context itself. It is currently known that factors such as the TATA-box and the sequences in promoter regions influence the expression variability, or in our case, the BCV$^2$. In order to fully understand the extent of the influence of the nucleotide sequence and surrounding genes on genetic variation, all genes after filtering were sorted into 5 bins sorted by the average expression ($\mu$) in fragments per kilobase per million reads (FPKM) and 5 bins based on their BCV$^2$.

We observed that a high GC% around the transcription start site (TSS) is associated with both a lower BCV$^2$ and lower mean expression level (figure 1 A-B). It should be noted that a higher GC% downstream of the TSS is more likely to be observed in genes within a quintile of highest expression variability, but the separation is less pronounced for genes with a high average expression ($\mu$) (Figure 1 A-B). The average expression and BCV$^2$ correlates with GC content at ~30-50 bp upstream of the TSS, which overlaps with position of a TATA-box (figure S2).

Predictive models for the BCV$^2$ and average expression based on the presence of different transcription factor binding sites (TFBS) have shown that the presence of a TATA-box was indeed the strongest predictor amongst all TFBS for both BCV$^2$ and the average expression (Figure S3). To identify the relation between DNA sequence of genes and the fluctuations in BCV$^2$, the melting temperatures for the gene nucleotide
sequences were calculated in a position specific manner. It was found that genes with a higher duplex stability had a larger variation in gene expression (Figure S14).

Figure 1. A) average coverage of GC% for genes separated into quintiles sorted by BCV². B) average coverage of GC% for genes separated into quintiles sorted by average expression. C) Overview of correlations between n2 position dependent models predicting BCV2 and average expression in FPKM and measured values. \( \lambda_{\text{min}} \) represents the model with the lowest error, though this might be overtrained. \( \lambda_{\text{se}} \) represents the simplest possible model which performs within one standard error of the optimal model. D) Gene-wise predicted BCV² plotted against the observed BCV².

Upon observing the predictive value of the GC content for the BCV² and average expression, a position dependent di-nucleotide model was created to model the effect
of the genomic sequence on BCV^2 and the average expression (Figure 3C). Interestingly, a large part of the BCV^2 could be reliably predicted based on merely the genomic sequence between 1 kb upstream and 2 kb downstream of the TSS. Noting only the genomic sequence 1 kb upstream of the TSS, up to 500 bp downstream of the TSS had a predictive value to the average expression level of a gene. The strong correlation between sequence context around TSS and the BCV^2 proves that a part of the inter-individual non-Poisson variability stems from “intrinsic sources”, or characteristic gene noise.

**Expression variability is linked to nonproductive accumulation of Pol II**

The correlations between the genetic sequence and the BCV^2 mean that these impacts either the synthesis or degradation of RNA in order to be observed in RNA-sequencing samples. For this reason, we hypothesized that the higher duplex stability of the DNA could impact the speed or efficiency of RNA-Polymerase II (Pol II) during transcription. In order to investigate these, data from several datasets on the positioning of Pol II were correlated to the average expression and expression noise.

To detect the average occupancy and activity of Pol II on genes for which the average expression and BCV^2 were calculated we sampled different sets of sequencing data of genomic sequences bound by various forms of Pol II (Pol II, Pol II S2p and Pol II S5p) taken from mouse liver samples from public databases (Koike et al., 2012).

We found that genes with the highest BCV^2 were also more often occupied by Pol II S2p, implying that genes which are highly variable in expression have more Pol II bound to their gene bodies (figure 2A). Similarly, genes with a higher average expression have a higher occupation of Pol II S2p on the gene body (figure 2B). This is an interesting observation, as we observe that the BCV^2 is not correlated to the average expression (Figure S1D). This must mean that genes which have a higher occupation of Pol II, can either be more variable in expression, have a higher expression, or both.
though the two observations are unrelated. This observation was verified with all forms of Pol II, Pol II S2p and Pol II S5p (Figure S5).

Despite the knowledge of an increased binding of Pol II among genes with a higher average expression and variability, it is not known whether this accumulation results in nascent transcripts. Global Run On sequencing (GRO-seq) data allows for the retrieval of such data, revealing the genomic location and amount of recently produced RNA in a sample (Core, Waterfall, & Lis, 2008).

Figure 2. A, B) Average coverage of Pol II S2p binding separated per quintile based on BCV² and average expression in FPKM. C, D) Average coverage of nascent transcripts as a result of GRO-seq data separated per quintile based on BCV² and average expression in FPKM.

The GRO-seq reads obtained for mouse liver samples (Fang et al., 2014), were aligned and quantified. The genes were then grouped both by BCV² and mean expression from the diversity outbred dataset (Figure 2C-D). An apparent stratification of the quintiles can be observed between GRO-seq reads and the average expression of genes (Figure 2D), but no such separation is visible for the BCV² (Figure 2C). This implies the act of
transcription, as illustrated by GRO-seq data, does not influence expression noise, rather the speed or efficiency of said transcription, which is illustrated by the relative over-abundance of Pol II occupation, is contributing to the overall expression noise \( (BCV^2) \). This means that the increased Pol II occupation in genes with a high variability in expression is non-productive and does not result in an increased number of nascent transcripts as evidenced by GRO-seq data.

**Epigenetic determinants of transcription noise**

Observing the differences in \( BCV^2 \) between genomic regions with a high GC content, and the subsequent non-productive accumulation of Pol II lead us to investigate the characteristic epigenetic profiles associated with variably and robustly expressed genes. Epigenetic marks could cause the non-productive accumulation of Pol II or could be disrupted by the accumulation itself.

Whole Genome Bisulfite Sequencing (WGBS) data revealed that a higher methylation around the TSS more often coincided with a higher \( BCV^2 \), whilst further into the gene body this effect becomes inverted, where it seems a high methylation of the DNA is correlated to a lower \( BCV^2 \) (Figure 4A). This observation correlates with the observation that genes with higher levels of methylation in their promoter region have a higher complexity of initiation, influencing \( K_{on} \) and \( K_{off} \) ratios.

Genes with a high average expression have a clear lack of methylation on the gene body, as the 20% of genes with the highest average expression have the lowest methylation levels in the gene body (Figure 3A). Both the genes with the highest 20% \( BCV^2 \) and the highest 20% in average expression were observed to have fewer CpG sites around the TSS (Figure S6). Methylation of cytosines in the region around the TSS implies the unavailability of a gene for transcription. To further delve into relation between expression variability and the availability of the TSS and the surrounding area DNAse-seq data was utilized to identify the regions with open chromatin and its possible correlation to the \( BCV^2 \) and average expression.
Reads were again mapped, quantified and coverage was shown for all genes separated by either BCV² or average expression. It was found that a higher accessibility of the DNA correlated to both a higher BCV² and average gene expression (Figure 3B).

**Figure 3.** A) Average coverage of meCpG WGBS reads separated per quintile based on BCV² and average expression in FPKM. B) Average coverage of DNase-seq reads separated per quintile based on BCV² and average expression in FPKM.

Formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) data verified that genes which are free from protein (transcription factors, histones, etc.) around the TSS are more likely to have lower BCV², whilst genes with a higher average expression tend to have a higher frequency of proteins bound to the DNA around the TSS (Figure S7).
Though FAIRE-seq data revealed that genes which have less proteins bound downstream of the TSS are more likely to have lower BCV². This observation implies that there might be significant differences in the density and spacing of histones for genes with different expression variability.

**Nucleosome positioning**

The reduced density of proteins bound downstream of the TSS implies that the nucleosome binding, at and downstream of the TSS might be reduced for genes with higher BCV². To investigate this, publicly available MNase-seq data of mouse liver cells under normal conditions was taken from GEO (Menet, Pescatore, & Rosbash, 2014), were plotted, and were again separated according with the BCV² and FPKM quintiles (Figure 4). This showed that not just the availability of the TSS was higher on average, but also the spacing of the nucleosomes was less regular for genes which fell into higher quintiles separated by BCV² (Figure 4A). No such separation was observed when explore the nucleosome phasing based on the average level of gene expression (Figure 4A).

By creating a predictive model for the BCV² and average expression based on the nucleosome occupancy per gene, the predictive value (β) of a position occupied by a nucleosome could be calculated. This model verified that if a position is occupied by a nucleosome the predictability of a gene’s BCV² significantly increases, whilst the average expression in FPKM does not (Figure S8A). Fourier transformation on the phasing of the nucleosome occupancy found that genes with lower BCV² had less variation in the distance between nucleosomes than genes that displayed a high BCV² (Figure S8 B). Variation in this distance did not have any predictive power for the average expression, except for genes within the quintile with highest expression, for which the nucleosome phasing was less strict than that for the genes within the lower expression quintiles.

This comparison was repeated with an independent dataset (GSE57559), to prove consistency across different experiments and even with different concentrations of
MNase. We observed the same trends for both the BCV² and the average expression levels (Figure S8C).

It is known that the chromatin structure is modulated by histone modifications which impact the nucleosome structure and positioning (P. Zhang, Torres, Liu, Liu, & Pollock, 2016). Observing a change in the regularity of the space between the nucleosomes we wondered if certain histone marks were associated with a higher average expression or BCV². ChIP-seq data taken from ENCODE (Li et al., 2014) was aligned, quantified and the coverage for each gene was again grouped and sorted per quintile by the average expression (FPKM) and noise (BCV²) of DO mice under standard chow diet (Figure S9).

We found that not only do certain histone marks occur more often at specific positions after the TSS, but that often the presence of histone marks has a predictive value for both the average expression (FPKM) and the noise (BCV²). We created a predictive model for the average expression (FPKM) and the BCV² based on several methylation and acetylation marks (figure 4b) and found a high predictive value for both the BCV² and the average expression. Notably, the position-dependent predictive value of histone marks on the BCV² is much higher for regions after the TSS, meaning the impact of histone marks is greater if these inhibit transcription elongation rather than transcription initiation.

**Diet can modulate expression variability**

Variation in gene expression is exemplified by the distribution of expression levels among samples, which follows a mixed Poisson distribution. The non-Poisson part of the distribution finds its source in both characteristic gene noise and system state variables. As we discussed above, we observed several intrinsic factors (genetic and epigenetic) acting as upstream cellular drives impacting the BCV². Because all mice were held under similar laboratory conditions and all samples were taken from the identical organs, no effect of system state variables could be identified. When
analysing samples from mice kept under different conditions it is possible to identify the impact of system state variables on both the overall, and gene-wise $BCV^2$.

**Figure 4.** A) Z-scores of expected/observed nucleosome bound reads separated per quintile based on $BCV^2$ and average expression in FPKM. B) Correlation of predicted $BCV^2$ and predicted average expression in FPKM based on a multitude of histone marks for both $BCV^2$ (left) and average expression in FPKM (right).

Comparisons of genes expressed constitutively in liver samples of diversity outbred (DO) mice fed either a standard (Std) or high fat (HF) diet revealed an overall increase in $BCV^2$ under HF dietary conditions (Figure 5A).
However, due to the way BCV² is calculated from RNA-seq data its value will be dictated by both the characteristic gene noise and system state variables. Therefore, the identification of a system state variable allows for the calculation of the impact of the factor on the BCV².

![Figure 5](image)

**Figure 5.** A) boxplots of BCV² for mice fed a standard (std) or high fat (HF) diet, separated by all, male and female samples. B) boxplots of corrected BCV² for mice fed a standard (std) or high fat (HF) diet, separated by all, male and female samples. C) Overview of log fold changes in BCV² from a Std to a HF diet (left panel), calculation of the fold change in system state (middle), and the fold changes after correction for the change in system state (right).

In this case, because the system state variable *Diet* is known, it is possible to calculate the slope of the standardized major axis between the BCV² for standard diet and high fat diet through a method similar to principal component analysis approach (supplemental equation 18-19).
This will allow for the identification of the change in expression variability ($BCV^2$) due to gene-specific characteristics. Removing the system state variable from the equation allows us to observe whether the change in $BCV^2$ under differing dietary conditions affects the system state part of the noise or the characteristic gene noise. In other words, by identifying the impact of diet on $BCV^2$, we can calculate a $BCV^2$ adjusted for diet (Figure 5B).

Investigating the fold change in $BCV^2$ between a high fat diet for all samples, males and females (Figure 5C) we observe that the $BCV^2$ clearly increases due to both system state variables (Diet) and changes in the characteristic gene noise.

Calculating the fold change for the system state by identifying the slope of the major axis we observe that a high fat diet has a strong impact the overall $BCV^2$ (Figure 5C). After the correction for diet as a system state variable the noise ($BCV^2$) is still significantly increased between all samples, males and females (Figure 5B). This means that the change in diet does not only affect the system state component of the noise, but also the characteristic gene noise. This means that factors, such as the synthesis rate, burst size and/or burst frequency are affected at different rates among cells as result of the change in diet.

Investigation of the $BCV^2$ in inbred mice strains revealed a more varied response, indicating that the overall increase in $BCV^2$ due to a HF diet is strain-specific (Figure 6). Expression variability in inbred mice both under a standard and high fat diet shows a strong correlation in $BCV^2$ between all samples, indicating that a large portion of the noise ($BCV^2$) is shared (Figure S2).

Though the greater variation estimates could relatively low number of replicates (8 biological replicates per inbred strain). We calculated the standardized major axis (SMA) slope between the $BCV^2_{hf}$ and $BCV^2_{sd}$ resulting in an adjusted $BCV^2$ for samples under differing dietary conditions. The New Zealand Obese (NZO), Watkins Star Line B (WSB) and Black Six (B6) had an overall reduction of $BCV^2$, yet after correction for
high fat as a system state parameter these show the largest increase in corrected BCV² (Figure 6).

This implies that expression variation (BCV²) in these samples is strongly affected by the system state parameter, rather than the characteristic gene noise. The other breeds (B6, NZO and WSB), with the exception of the Non-Obese Diabetic (NOD) strain showed an initial increase in BCV², yet this effect is mostly connected to the system state change brought on by the difference in diet.

Most inbred mouse strains used in this investigation have a medium or low susceptibility to obesity. The mice strains derived from wild isolates (CAST, WSB, PWK) are generally among the leanest. B6 being often used for dietary experiments due to its susceptibility to gain weight, and NOD being shown to be slightly heavier than the average mouse in other investigations (Reed, Bachmanov, & Tordoff, n.d.). The clear exception is the New Zealand Obese (NZO) mouse, which has a high susceptibility to obesity (Ortlepp et al., 2000). So far, there is no direct correlation between the weight of mice, yet the differences in response to a high fat diet in terms of noise (BCV²) suggests a strain-dependent response to system state variables rather than a universal mechanism which dictates expression variability (BCV²).
**Gene expression variability characteristics are similar in mice and rats**

Though we observed that the GC-richness and the di-nucleotide content were predictive for the BCV² among mice, the impact of these factors might just correlate with other gene characteristics. To verify whether the genetic sequence does dictate variation in gene expression across varying expression levels, organs and species we utilized publicly available RNA-seq data taken from rats (Yu et al., 2014). This data was sampled for rats of both sexes, from 11 different organs, at 4 different ages, resulting in a total of 320 samples. Normalization was performed as was with the mouse RNA-seq data and, similarly to mouse data, no correlation was found between the average expression and BCV² (Figure S11). A predictive model based on KEGG pathways for the BCV² across all samples again showed that genes associated with processing of genetic information processing are generally expressed at more robust levels, whilst genes associated with metabolic pathways (PPAR signalling) were more predictive for a higher variation in expression (Figure S12). Comparison of the BCV² across samples and per organ showed correlations across all groups (Figure 7), illustrating that the BCV² of a gene is conserved across organs. Moreover, comparison of homologues genes between mice and rats shows that BCV² values exhibit high interspecific correlation (Figure 7B).

Separating the GC% of genes per quintile based on the BCV² across all organs for rat samples resulted in a similar laddered result that was also observed in mouse samples (Figure 8A). A higher GC% downstream of the TSS is more common in genes with a high BCV² Consequently, calculation of the duplex stability of expressed genes shows a similar correlation between a high energy cost of transcriptional activities and the average BCV² as seen in the mouse RNA-sequencing data (Figure S13).
A second position-dependent di-nucleotide-model was created to predict the BCV$^2$ of rat genes, which was simultaneously tested on mouse genes to see how well the model would carry over across species (Figure 8b). Interestingly, the results of BCV$^2$ modelling indicate that similar regions are important for predicting the BCV$^2$ of mice and rats. Further, a similar correlation was observed when predicting BCV$^2$ for rat genes with a model trained on mouse data (Figure 8c).

The predictive power across species can be further increased by the implementation of a non-position specific hexa-nucleotide predictive model (Figure 8d). Surprisingly, this model does not improve the predictive capability within species but does increase significantly between species.
The predicted BCV² for mice based on a rat model has a correlation only two points below the actual correlation between the BCV² of mice and orthologous rat genes (Figure 8d). In terms of R² this means the model solely based on the genetic sequence is almost as accurate as using the orthologous gene as a reference.

**Gene expression variability changes with age**

Comparison of the BCV² between samples of different ages shows a consistent increase of BCV² with age (Figure S14). Grouping of the samples into organ-specific bins mostly maintains this linear increase of BCV² with age, though it does not hold true across all organs. However, the fine binning of the data creates smaller groups, which are more
susceptible to errors in variability estimates. The correlations in variation estimates (BCV^2) across different ages, organs and sex is exemplary for the impact of both system state parameters and characteristic gene noise. A part of the variation stems from systematic factors which affect all cells within a sample independently, this includes organ, age and sex, another part stems from the characteristic gene noise, which is dependent on the genetic sequence and internal factors which affect the RNA synthesis or degradation rate.

The separation of the gene expression variability observed in RNA sequencing data in Poisson and non-Poisson parts allowed for the complete removal of the correlation between the average expression and the variation in the form of the BCV^2. With the identification of the fluctuation response relationship within bulk RNA-seq data we showed that cell-to-cell variation is carried over to inter-individual noise level. This became abundantly clear upon the creation of di-nucleotide models for the prediction of variability, which showed a large portion of the variation in gene expression stems from not only the composition of the promoter region, but the region downstream of the TSS too. This increased variability strongly correlated with the non-productive accumulation of nucleosomes, which co-occurred with histone modifications. Not only do these sources of characteristic gene noise influence the overall BCV^2 of a gene, so do external factor such as diet, organ and age. These observations will be of great importance to the field as these lay the ground works of the intricate factors which together dictate the variation in gene expression.

**DISCUSSION**

One of the most commonly mentioned observation within the field of expression noise is that the coefficient of variation (CV) is dependent on the average expression. By showing that with an average expression as a result of a synthesis/degradation process there will always be a Poisson component to the distribution, and subsequently subtracting said component we have defined noise as a doubly stochastic process which stems from both the Poisson and the non-Poisson
components. The non-Poisson noise is caused by fluctuations in the synthesis rate, degradation rate or $K_{on}/K_{off}$ ratios, or in other terms fluctuations to the burst size and frequency.

It is often assumed that cell-to-cell variation in gene expression is lost in bulk RNA-sequencing experiments. Logically this would make sense, as the RNA-sequencing multiple cells from a bulk sample results in a summation of reads in which all variation between cells is lost to a singular average. However, due to the fluctuation response relationship, which states the initial variation in expression impacts the increase in variation in expression due to external factors, the cell-to-cell variation is maintained in bulk RNA sequencing experiments. Without this equation, cell to cell variation could only be obtained from extensive single cell RNA-sequencing experiments, resulting in high costs and a large time investment. Now it would be possible to perform single-cell RNA sequencing experiments to capture variation caused by kinetic parameters in combination with bulk RNA-sequencing to capture the magnitude of the system state fluctuation.

In this chapter we elucidated the BCV$^2$ as a product of both the characteristic gene noise and system state parameters. The characteristic gene noise encompasses all internal factors which impact the variation in gene expression as would be typical for specific genes. This would include the promoter architecture, the methylation state and the burst frequency. The system state parameters are factors which alter the variation in gene expression between cells and individuals equally, these factors include diet, age, sex and organ. Due to this separation of system state and characteristic gene noise we obtain the possibility to separate the BCV$^2$ into the part which is caused by the known system state parameters and ‘the remaining noise’, which consists of both the characteristic gene noise and undefined system state parameters. This separation is highly valuable in future RNA-sequencing research as it opens a door to a highly under-explored portion of information which is contained within RNA-sequencing data.
In this investigation we showed that not only does the promoter region or the presence of a TATA-box influence the gene expression variability (BCV²), but the characteristic gene noise is strongly influenced by the genetic sequence downstream of the TSS. This observation holds true between different mouse strains, across species and different organs. Within the same species a hexanucleotide semi-position dependent model did not improve the prediction for the BCV² illustrating that a large portion of the variation is simply a consequence of the nucleotide composition of the gene, which in turn affects other factors which influence the transcription rates. The hexanucleotide semi-position dependent models did perform better across species, there could be several causes for this: 1) The system state variables for rats are vastly different than that of mice. This means physiologically rats are just very different from mice that is reflected in different gene activity, metabolic state, differences in circadian rhythm etc. This explains why within rat species the accuracy of predictions of BCV² are much higher, as the system state variables between these rats are much more identical. 2) The annotation of the rat genome is not as complete and precise as that of the mouse genome, meaning the annotated TSS and promoter region are not always right. Hence, the models trained on mouse data/genes would then not carry over that well to rat data/genes.

After establishing the predictive capabilities of the dinucleotide content of genes, we identified possible consequences of the sequence composition. We observed an accumulation of Pol II on the genes which under normal circumstances have either a high average expression or a high amount of variation in gene expression (BCV²). Though no increase in nascent transcripts was observed for genes with a high BCV². This implies that even though Pol II does bind more frequently to genes with a higher variation in expression, these polymerases do not finish transcription, or are “stuck in traffic” unlike genes with low variation in expression. Though we do now understand the increased frequency of non-productive Pol II on genes increases the variation in expression levels determined from RNA-sequencing data, we do not know whether the sequence, and thus the melting temperature of the DNA duplex is the main cause
of this. Another possibility is that the genetic sequence causes the preferential binding of proteins, which in turn functions as speedbumps for polymerase II as it attempts transcription of a gene.

In this chapter we did observe an increase of methylation of the TSS, a preference for certain histone marks, and a less regular phasing of nucleosomes on genes with a high variation in BCV². Predictive models of histone marks based on DNA sequence have been shown to work (Whitaker, Chen, & Wang, 2015), though we have not explored this correlation in this investigation.

Yet the increased occurrence of certain histone marks could be a consequence of the high variation in transcription instead of a cause. ChIP-seq experiments taken from public resources have allowed us to gain an overview of the nucleosome distribution and histone modifications that frequently occur in genes with a high BCV². We have showed that not only do certain histone marks co-occur more often in genes with a high BCV² in mice on a standard diet, but that the distribution of nucleosomes is less frequently phased among genes with a high BCV². Protein sub-complexes have been shown to exist which aid in the replacement of nucleosomes after Pol II has transcribed a section of a gene (Kulaeva, Gaykalova, & Studitsky, 2007; Kwak & Lis, 2013). With the understanding that the di-nucleotide content increases noise and reveals itself in the non-productive accumulation of Pol II in the gene body, it could be that the less strict organisation of the nucleosomes is a consequence of polymerase crowding, in which the high amount of Pol II does not leave enough space or time for the nucleosomes to be properly reinstalled after each passing of Pol II.

Besides the effects and sources of characteristic gene noise we also investigated several system state parameters. These are factors which impact all cells in a sample or organisms in a group equally. A simple example is the relative increase in BCV² among diversity outbred (DO) mice under a high fat diet (HF) as compared to mice on a standard diet (Std). It has long been known that dietary restriction can have positive effects on the lifespan of organisms (Lee & Longo, 2016). A recent study has shown that a mutation which mimics a dietary restriction which has both beneficial effects
on the life and health span of mice, reduces the overall expression variability too (Müller et al., 2018). In this investigation we observed that rats at a more advanced age showed an overall higher variation in gene expression ($BCV^2$). To our knowledge, there was no direct investigation of expression variability in which a combination of ageing and different diets have been performed as of yet, we expect that the increase of $BCV^2$ which occurs at an advanced age is reduced under a caloric restriction diet and is possibly aggravated by high fat or high caloric diets. Through correction for the known system state parameter, or diet, we were able to calculate a corrected $BCV^2$ for both DO and inbred mice strains, in which we observed that not only does the system state part of the noise change, so does the characteristic gene noise. This leads us to conclude that not only do system state parameters alter the $BCV^2$, these can directly impact the characteristic gene noise. In the case of an increased adjusted $BCV^2$ under a high fat diet there are several attributes of the characteristic gene noise which could be affected. The synthesis rate could accelerate due to a higher availability of energy, a higher production of polymerase II or a greater abundance of free nucleotides because of the richer diet. Another possible effect lies within the methylation which can be altered by a shift in diet (Y. Zhang & Kutateladze, 2018). The presence of alternative histone marks could interfere with the speed and efficiency of transcription, resulting in a higher variation in RNA abundance ($BCV^2$) from cell to cell.

Due to the better implementation of the mathematical models that underlie the variation in gene expression we suggest the new way for further research to elucidate the sources of variation in gene expression. The proof that the fluctuation response relationship enables us to estimate variation in gene expression that also occurs from cell to cell and among individuals can be of immense value, potentially reducing the need for extensive single cell sequencing experiments in the future. Our analysis allows for the separation of characteristic gene noise and the noise caused by known system state parameters, which influence each other. We found several sources underlying and co-occurring with an increased expression variation and have taken
several steps in the identification of factors that modulate gene expression variability. With the observation that variation in gene expression increases with an advanced age, and different factors can aid in the reduction of expression variation we have gotten one step closer to monitoring and prospectively counteracting the adverse effects which come with the ageing process.

MATERIALS AND METHODS

1086 RNA-sequencing libraries taken from mouse liver samples from diversity outbred and inbred mouse strains under differing dietary conditions (Standard chow, high fat) were retrieved from the Gene Expression Omnibus (Munger et al., 2014). STAR 2.5 was used to align the reads against GRCm38 and output to count tables. Read counts were upper quantile normalized and all genes with less than 1 count per million (CPM) for each sample were removed from the analysis. Average FPKM values were calculated for each test group (Standard chow, high fat), separately for each strain of mice. The biological coefficient of variation (BCV² / E-CV), was calculated in line with the methods presented in chapter 2.

RNA-sequencing libraries of Rattus norvegicus, taken from 10 different organs, at four different ages in both males and females to a total of 320 samples were retrieved from GEO (Yu et al., 2014). STAR 2.5 was used to align the reads against Rnor6.0 and output to count tables. Samples were grouped by sex, organ and age for separate tests. FPKM and BCV² values were calculated in the same manner as mice samples.

GAMLSS was used to calculate the biological coefficient of variation (BCV²) as shown in chapter 2 of this thesis (supplementary notes).

The technical coefficient of variation was calculated with GAMLSS as above with technical replicates included as a factor for overdispersion (α). A log-likelihood ratio test (LR-test) comparing Poisson and Negative-Binomial models was used to identify the significance of the overdispersion parameter. From this test we concluded that the
overdispersion due to technical replicates was insignificant and negligibly small, so no adjustment was made to the data for technical variation.

The free energy ($\Delta G$) for DNA/DNA and RNA/DNA duplex stabilities were calculated for all genes in kcal/mol per position relative to the TSS based on that estimated for dinucleotides (Tulpan, Andronescu, & Leger, 2010). The averages per position were plotted per FPKM and BCV quintiles for each gene. Combined energy costs for transcription sum from DNA/DNA $\Delta G$ required for dsDNA melting and RNA/DNA $\Delta G$ required for separation of newly synthesized RNA.

To predict coefficients of variation for genes $g_i$ from DNA sequence or chromatin context, or from the KEGG biological pathways we applied Ridge regression (Hastie, Tibshirani, & Friedman, 2001) implemented in R glmnet package (Friedman, Hastie, & Tibshirani, 2010). Models were trained with 10-fold cross validation. Penalty $\Lambda$ was chosen either from the best model with lowest cross-validation error - $\Lambda_{\text{min}}$ or from the simplest model with cross-validation error within one standard error of the best model - $\Lambda_{\text{se}}$. For most of the models, unless specified, we set $\Lambda$ at $\Lambda_{\text{se}}$. In training of models, we also normalized $\alpha_{g_i}$ (response vector) as we noted that models’ intercepts $\beta_0$ could not be predicted from the intrinsic parameters (DNA context, chromatin modification or KEGG pathway), but rather adjusted by extrinsic context (cell type, tissue, diet, age, etc.).

Similar ridge regression models were used to predict mean-variance normalized parameters of gene expression ($\text{BCV}^2$, FPKM) for mouse and rat liver samples based on position dependent di-nucleotide, tri-nucleotide and hexa-nucleotide content of all protein coding genes with at least 1 count in all samples. Accuracy of the models was validated across species.

A LR-test was used to identify genes that have a significant change in average expression (FPKM) or $\text{BCV}^2$ (FDR < 0.01). The $\text{BCV}^2$ and fraction of significantly different $\text{BCV}^2$ s were visualized in various ways with R.
All protein coding genes from mice liver samples under standard dietary conditions with more than 1 CPM in all samples were grouped into quintiles based on FPKM and BCV². For several characteristics (Pol II-seq, GRO-seq, DNAse-seq, etc.) the average coverage per position, up- and down-stream of the TSS for each factor was calculated per bin for both FPKM and BCV² bins. This coverage was in either percentage or Z-score based on all samples. The five groups were visualized in R over several ranges up- and down-stream of the TSS.

The presence of the TATA motif; was taken from the Jaspar database (Khan et al., 2018) for every position between 5kb up- and downstream of every gene within in the count tables. The relative frequency was calculated per quintile of FPKM and BCV for the diversity outbred set as shown before.

Chip seq occupancy profiles for the S2 phosphorylated (S2p) and S5 phosphorylated (S5p) RNA Pol II were retrieved from GEO (GSE41472) (Li et al., 2014). Pol II-seq reads were aligned to GRCm38 with STAR aligner and output to count tables. The libraries were upper-quantile normalized. The average coverage per base pair was calculated 5kb up- and down-stream of the TSS and TTS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of Pol II occupancy per quintile based on the average expression in FPKM and BCV for the diversity outbred set. The results were plotted with R.

GRO-seq transcripts taken from mouse liver samples taken during different stages during light/dark cycle were retrieved from GEO (GSE59486) (Fang et al., 2014). GRO-seq reads were aligned to GRCm38 with STAR aligner and output to count tables. The libraries were upper-quantile normalized. The average coverage per base pair was calculated 5kb up- and down-stream of the TSS and TTS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of GRO-seq read coverage per quintile based on the average expression in FPKM and BCV for the diversity outbred set. The results were plotted with R.
WGBS reads were aligned to GRCm38 with STAR aligner. The libraries were upper-quantile normalized. The average coverage per base pair was calculated 5kb up- and down-stream of the TSS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of WGBS read coverage per quintile based on the average expression in FPKM and BCV for the diversity outbred set. The results were plotted with R.

DNAse treated genomic sequencing data taken from 8 week old adult mouse (C57 black 6) liver samples were downloaded from GEO (GSM1014195) (Vierstra et al., 2014).

DNAse treated genomic sequencing reads were aligned to GRCm38 with STAR aligner. The libraries were upper-quantile normalized. The average coverage per base pair was calculated 5kb up- and down-stream of the TSS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of DNAse read coverage per quintile based on the average expression in FPKM and BCV for the diversity outbred set.

Dyads of nucleosomes were calculated based on 24/47 Mnase-Seq libraries retrieved from GEO (GSE47142) (Menet et al., 2014).

The histone bound genomic reads were aligned to GRCm38 with STAR aligner. The libraries were upper-quantile normalized. The middle of the 150bp reads (dyads) were mapped 5kb up- and down-stream of the TSS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of read coverage per quintile based on the average expression in FPKM and BCV for the diversity outbred set and were plotted in R.

12 MNase-seq libraries taken from adult mice livers were taken from GEO, (GSE57559) (Iwafuchi-Doi et al., 2016).
H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2 sequencing data was aligned with star and matched against 5 kb region up- and down-stream of the TSS of known protein coding genes. The libraries were upper quantile normalized. The genes were then sorted into quintiles based on the BCV and the average FPKM. Two separate linear models were created based on the BCV and the FPKM of genes of the different histone modification sequencing data. The resulting model was then correlated to the actual sequencing data per 1000 bp bins with no overlap, relative to the TSS.

FAIRE-seq data taken from livers of C57BL/6J and A/J mice on three diets was taken from GEO, (GSE75984) (Leung, Trac, Du, Natarajan, & Schones, 2016). FAIRE-seq sequencing reads were aligned to GRCm38 with STAR aligner. The libraries were upper-quantile normalized. The average coverage per base pair was calculated 5kb up- and down-stream of the TSS of known protein coding genes with a minimum expression of 1 CPM. Z-scores were calculated to show the relative abundance of FAIRE read coverage per quintile based on the average expression in FPKM and BCV for the diversity outbred set.

REFERENCES


Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., ... Mathelier, A. 202


RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics, 198*(1), 59–73. https://doi.org/10.1534/genetics.114.165886


of promoters on noise in gene expression using thousands of designed sequences. *Genome Research, 24*(10), 1698–1706. https://doi.org/10.1101/gr.168773.113


expression level and noise by histone modifications. *PLOS Computational Biology, 13*(6), e1005585. https://doi.org/10.1371/journal.pcbi.1005585


SUPPLEMENTARY MATERIAL

In order to lower the total number of pages, all supplemental material of this thesis has been moved online. For additional reading, please see the following link:

https://drive.google.com/open?id=1skLP9E2hToDLGgkNrmCoWcTVeBmiZX1-

If you might encounter a broken link, please contact me:

Tristan_dejong@hotmail.com