



UvA-DARE (Digital Academic Repository)

Understanding gene expression variability in its biological context using theoretical and experimental analyses of single cells

Kempe, H.

Publication date

2017

Document Version

Other version

License

Other

[Link to publication](#)

Citation for published version (APA):

Kempe, H. (2017). *Understanding gene expression variability in its biological context using theoretical and experimental analyses of single cells*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Chapter 1

Introduction

Abstract

Living organisms are constantly exposed to fluctuating environments. The inherent variability of cells is an important mechanism for cell survival under fluctuating conditions. However, cells are also able to respond to environmental changes by cellular adaptation through complex regulatory mechanisms. Multicellular organisms consist of a variety of cell types that exhibit unique cell identities. Although all cells of an organism possess the same genetic material, the conversion of this information to functional molecules (i.e. gene expression) is regulated in a cell type-specific manner. Epigenetic gene regulation confers stability of cell type-specific gene expression patterns while allowing changes in response to environmental or developmental cues. Typically, epigenetic regulation causes the phenomenon of inheritance of a property that is not encoded in the DNA base pair code. Gene expression cell-to-cell variability, or 'noise', can also determine important cellular processes such as development and differentiation but it can also perturb cellular regulation leading to disease development or altered drug responsiveness. Gene expression variability is affected by the dynamic nature of cells, for instance via cell cycling and division events. During these processes, the intracellular composition of a cell changes constantly and leads to variability in gene expression. Cell-to-cell variability is largely affected by biochemical noise inflicted by the random diffusion and reaction probabilities of molecules. Understanding molecular principles of biological variability in multicellular organisms is a major focus of fundamental life sciences aiming to understand healthy developmental processes involving cellular (re)programming. Moreover, such knowledge largely impacts medical research towards understanding the development of diseases. In this thesis, we studied the impact of cell growth, an inherent property of living cells, on gene expression variability. Additionally, we studied gene expression cell-cell variability upon inducing an altered cellular state by (i) targeting an epigenetic regulatory protein to a defined reporter gene thereby changing chromatin composition and regulation of the reporter gene, (ii) exposing cells to UV-C thereby inducing DNA damage and (iii) providing long-term breast cancer (endocrine) treatment to follow the development of treatment resistance. The work presented in this thesis combines experiments, theory, and simulation of gene expression regulation to understand gene expression cell-cell variability.

1.1 Gene expression regulation

DNA is the genetic material that resides in the cell nucleus. Genes are genetic regions that provide the information to produce proteins. Gene expression represents the transfer of genetic information into functional molecules. The protein coding genetic information stored in DNA is transferred to RNA and subsequently to amino acid sequences in proteins. [1]. This sequential process can be split in two separate processes: gene transcription, which refers to the production of RNA, and translation, which refers to the production of proteins (Figure 1.1).

At the DNA level, transcription of DNA into mRNA depends on transcription factor binding at the non-transcribed region of a gene. Transcription factors can both activate or repress their target gene. Bound transcription factors can mediate binding of RNA polymerase II (RNA Pol II) together forming the pre-initiation complex. This complex can cleave DNA strands in order to start transcription elongation. During transcription elongation, RNA Pol II molecules slide over the DNA strand, read the genetic information, and synthesize RNA molecules. In eukaryotes, the RNA molecule produced during transcription of coding regions is subject to additional editing (addition of a 5' cap, polyadenylation, and splicing), which transforms RNA into messenger RNA (mRNA). After translocation of mRNA to the cytoplasm, mRNA can be translated by ribosomes into functional proteins. Transcription initiation can be induced via enhancers or cis regulatory non-coding DNA sequences proximal or distal from the regulated gene (5' upstream of the promoter, in trans of the regulated gene or of neighboring genes or intergenic regions far away from the regulated gene (up to 1 Mbp)). Enhancer looping with the promoter of the respective gene enables interaction between transcription factors bound at enhancer regions and the RNA Pol II transcription machinery. The traditional view of gene expression does not consider perspectives on gene expression in which for instance noncoding RNAs are transcribed that interfere with transcription or translation [2, 3].

1.2 Phenotypic plasticity

All single cells in a multicellular organism exhibit the same genetic material (i.e. these cells are isogenic) but vary considerably in their phenotype and function. This variability is inherently related to variability in the conversion of genetic information into functional molecules providing cells with a different cell identity. Changes in gene expression allow cells to differentiate and react to changing environmental conditions in time and space. This regulated variability between isogenic cells within a population or organism is referred to as phenotypic plasticity [4, 5]. Phenotypic plasticity enables cells to change their phenotype depending on external stimuli. In single cell organisms, phenotypic plasticity is a way to cope with altered environmental conditions, whereas in multicellular organisms, phenotypic plasticity enables cells to obtain a defined cell identity or to form specific cell subpopulations (e.g. cancer stem cells). For instance, alterations in the epigenetic chromatin composition and associated changes in gene regulatory principles, i.e. network motifs providing feedback control, bistability, delayed responsiveness, memory and robustness are among the driving forces of phenotypic plasticity.

1.3 The epigenetic chromatin state

Gene activity is influenced by the chromatin composition in which the gene is embedded. Epigenetic gene regulation includes several hierarchical layers of regulation, e.g. nuclear positioning

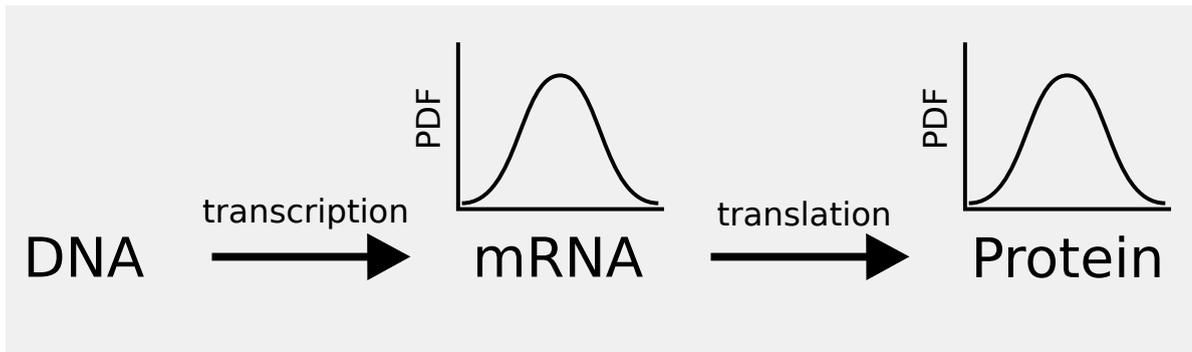


FIGURE 1.1: **The information flow from DNA to protein.** During gene expression, the genetic information stored in the DNA is transferred to mRNA and functional proteins. Both processes generate expression variability which leads to cellular variability in mRNA and protein expression.

of large-scale chromatin domains or gene loci, chromosomal interactions of for instance topological associated domains (TADs), conformational chromatin changes, nucleosome turnover, posttranslational histone modifications and the DNA methylation composition. Such epigenetic regulation largely determines gene expression patterning. The epigenetic state provides defined chromatin contexts that determine whether the chromatin stays in a permissive or restrictive state and whether genes become sensitive to transcription factors, epigenetic modifying enzymes and other regulatory proteins. The epigenetic chromatin state enables to establish and maintain cell type-specific gene expression. Nowadays it is clearly recognized that the local epigenetic composition of genes is involved in stochastic fluctuations in transcription. The stochastic conversion of the epigenetic chromatin state is still largely unresolved, specifically due to the lack of technologies to map the epigenetic composition in single cells. To fully recognize the contribution of the dynamics of switching between epigenetic chromatin states on the transcription output, it is essential to determine the epigenetic and transcriptional activity state in parallel in single cells.

1.4 Network motifs providing phenotypic plasticity

Expression is impacted by changes in gene regulatory networks. These gene regulatory networks can provide feedback control, bistability, delayed responsiveness, memory and robustness. Some network motifs enable a system to include memory to lock cells in certain expression patterns [6–8]. Here we focus briefly on two mechanisms that can provide memory of the transcriptional state and we show how this memory contributes to phenotypic plasticity.

Genes that stimulate their own production are subject to positive autoregulation. Positive gene autoregulation can lead to a bimodal distribution of the mRNA and protein levels of the

expressed gene within a population of cells. Positive autoregulation can induce the development of two cell populations, a subpopulation expressing low protein levels and a population expressing high protein levels. In Figure 1.2A, the network behavior of positive autoregulation is illustrated. For a cell to switch from a high to low expressing state, an external signal has to (temporally) stimulate expression, after which the protein levels remain altered due to the positive feedback [6, 7]. When the external signal, i.e. an environmental condition or fluctuation in regulatory components, is strong enough ($= 3$) the system adapts into a new steady state. Note that the conditions when no signal is present are the same before and after the stimulus, indicating the plasticity of gene expression. In engineered cell systems such regulatory networks are able to display predicted behavior [9–11] and are shown to drive defined cell behavior [12].

Double negative autoregulation is another network motif that is frequently occurring in gene expression networks (Figure 1.2B). This network motif consists of two gene products that mutually repress the expression of the other gene, thereby producing a switch-like gene network behavior [6]. An environmental condition or fluctuation (i.e. the signal) can temporally impact the expression or degradation of the involved genes. In this example, the altered condition induces degradation of gene product Y thereby releasing the repression on the other gene product, i.e. X . Upon removal of the environmental condition or fluctuation, X represses the expression of Y . Hence, the steady state expression levels in this gene network can be reverted under the same conditions. Experimental work of Gardner et al. showed that such a network leads to bistability in gene expression [10].

Positive and double negative autoregulation are two gene network motifs that are a few of the many naturally occurring network structures in cells [6]. Such network motifs illustrate that the previous (i.e. history) state of cells largely affects their cellular expression state thereby impacting cell-to-cell variability in gene expression

1.5 Variability in isogenic populations

Besides phenotypic plasticity, single isogenic cells within the same environment and with the same historical conditions can still display considerable variability in their gene production and hence phenotypic behavior. Isogenic variability can originate from intrinsic noise of biochemical processes, like random diffusion of molecules and reaction probabilities [13, 14]. Another important component of isogenic variability is the amplification of biological noise through regulatory cascades [13, 14].

One of the earliest notions that isogenic cells exhibit different phenotypic responses was documented by Bigger in 1944 [15]. A small subpopulation of *Streptococcus pyrogenes* bacterial cells survived penicillin treatment. However, upon an additional treatment a similar small subpopulation survived. With the use of single cell analysis techniques the molecular basis of such

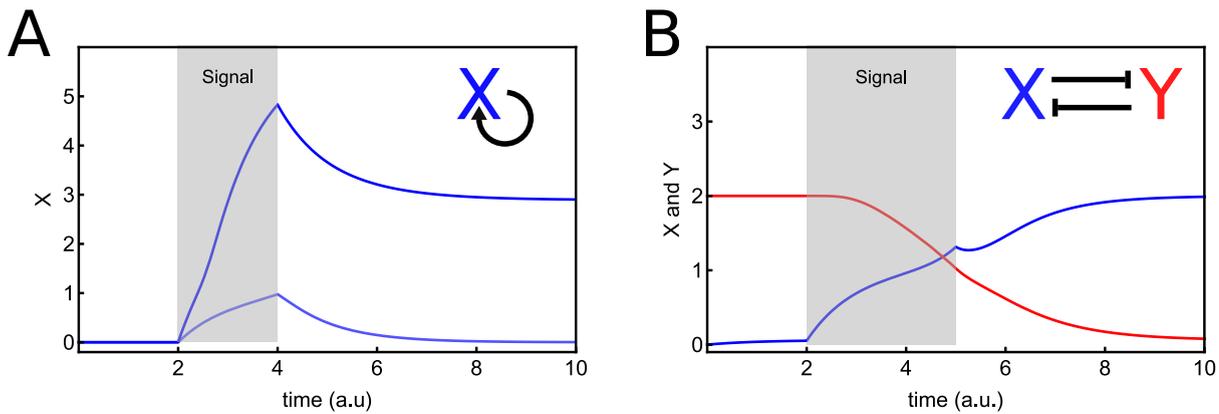


FIGURE 1.2: The cellular history can determine the transcription state of cells. A) A gene network motif representing positive autoregulation. A signal stimulates transcription of factor X. Factor X is produced with delay. The grey area illustrates the time the signal is present. The transcription of X remains elevated after the signal is terminated. The two blue lines represent the amount of X upon receiving a stimulating signal with high and low signal strengths. Positive autoregulation provides transcription regulation with delay and acts as a noise filter depending on the signal strength. B) A gene network motif representing double negative regulation. After a temporary signal (highlighted in grey), the expression states of X and Y are reversed. Double negative regulation enables to provide a stable active or repressed state that persists the absence of the initial signal.

population behavior could be identified. It was noted that cell survival is induced by unregulated fluctuations in the toxin/antitoxin system [16, 17], clearly indicating the benefit of cell-to-cell variability to enable cells to stay sensitive to penicillin treatment. Later studies showed that single cell variability in protein expression levels exist. It was shown that *Escherichia coli* defined cell populations consist of individual bacteria that either make β -galactosidase at a maximum rate or not at all [18, 19]. In 1997, McAdams and Arkin showed that gene expression regulation can account for expression variability within isogenic cell populations [20, 21]. Similarly, Thattai et al. [22] showed that the intrinsic variability in gene expression networks is larger during a transient reaction phase than in steady state. Combined, these studies indicate that gene regulation and dynamics can influence noise levels substantially. Interestingly, at that time experimental evidence for gene expression variability was not well developed.

Theoretical considerations of gene expression variability led to developments of novel experimental techniques to determine biological cell-to-cell variability. The use of fluorescent reporter constructs improved the ability to study single cell behaviour using fluorescent microscopy and flow cytometry and to infer cause-and-effect chains in gene expression cell-to-cell variability.

In 2002, Elowitz et al. [23] and Swain et al. [24] used *E. coli* cells expressing two fluorescently-tagged reporter constructs driven by the same promoter to show expression variability in a population of cells. Their experimental setup with two reporter constructs under control of the same promoter enabled them to distinguish two sources of gene expression noise, namely intrinsic and extrinsic noise. The definition of intrinsic and extrinsic noise depends on the description of

'the system' and 'the environment'. Elowitz et al. [23] used the following definitions. Intrinsic noise represents the expression variability within a single cell, and is quantified by the difference in expression levels of the two similar reporter constructs. Intrinsic noise is independent of the status of the host cell and it is especially large when the molecule numbers involved are low. In such instances, the collision events of randomly diffusing molecules are largely stochastic which contributes to intrinsic expression noise.

Extrinsic noise represents variability between cells in the population. Extrinsic noise is dependent on the status of the host cell and can be caused by cell volume growth, cell division events or variability in the concentration of involved expression regulators. A general phenomenon of extrinsic noise is that it can propagate through molecular networks within cells [25]. Raser et al. [26] showed that in eukaryotic cells extrinsic noise in gene expression is much larger than intrinsic noise. In their study, intrinsic noise represented at max 20% of the total noise. They concluded that variability in gene expression is not always reflected by the gene expression variability as measured with single-cell-reporter techniques and that extrinsic noise includes variability of the cell environment.

1.6 Constitutive and bursting transcription models

Technical advancements enabled measuring gene expression noise at the mRNA level. In contrast to the arbitrary protein readout units, mRNA is quantified in molecule numbers and can even be followed in real-time [insert reference]. mRNA counting accelerated the use of quantitative models such as the constitutive (birth-death) and bursting (two-state) gene expression models. Golding et al. [27] and Chubb et al. [28] showed that mRNA transcription proceeds in bursts, i.e. short periods of expression alternated with periods without expression. Transcription bursting implies that a gene can switch between an active and an inactive state. A combined experimental and modeling approach by Raj et al [29] showed that when the level of transcription factors increases, the burst size rather than the frequency increases. Additionally, the authors showed that the timing of transcription bursts is not controlled by a global factor. The timing of transcription bursts of two genes, integrated at the same genomic location, appeared to be largely correlated whereas transcription bursting of genes from different chromosomal positions is not correlated. This study indicates that genome position and epigenetic environment largely determine transcription bursting.

It has been hypothesized that transcription bursts are the result of changes in chromatin remodeling and in (un)binding of transcription factors. Vinuelas et al. showed that chromatin dynamics can influence variability in gene expression [30]. They created isogenic cell lines with a single fluorescent reporter gene integrated at different genomic locations (chromatin contexts). The expression of these reporter genes varied in their mean and variability level. They also

showed that expression variability is altered upon inducing changes in the chromatin structure by chromatin modifying agents. Their data indicate that the chromatin composition can contribute to gene expression noise.

1.7 Measuring expression variability

Gene expression noise can be measured at the level of mRNA and protein molecules. When gene expression variability is quantified at mRNA level, noise is caused by all processes that contribute to the production of transcripts. When gene expression variability is measured at protein level both transcription and translation contribute to the observed variability. The earliest experimental data on cell-to-cell variability in protein levels originates from bacteria and yeast cells transfected with fluorescent proteins like GFP, CFP, and YFP [23, 31–33]. Using fluorescent microscopy or fluorescence-activated cell sorting (FACS) single-cell protein profiles were measured. Such data can be represented as the amount of protein in a cell (in arbitrary units) or as the protein concentration (fluorescence corrected for cell size).

The introduction of experimental single-mRNA-sensitive detection methods opened up new possibilities to quantify gene expression noise at transcript level. The main methods used to quantify transcription noise are MS2/PP7-tagging and single-molecule mRNA FISH. The MS2-tagging technique exploits the binding capacity of fluorescently tagged MS2 bacteriophage-derived protein to MS2 stem-loop mRNA structures [34, 35]. MS2 stem-loop structures are incorporated in the untranslated region of the analyzed gene thereby enabling quantification of single fluorescently tagged MS2 transcripts in real-time. A major advantage of the MS2 method is that transcripts can be monitored in real-time in living cells. A disadvantage is that MS2-tagging requires genetic engineering.

Single-molecule mRNA fluorescence in situ hybridization (smFISH) is a technique to visualize endogenous mRNA not requiring genetic engineering [29]. Complementary DNA oligonucleotides are hybridized to the mRNA of interest. The mRNA is visualized via coupling of the oligonucleotides with a fluorescent label. The cells need to be fixed and permeabilised to enable binding of the probes. Hence, this technique provides transcript measurements of a single fixed time-point. The detection accuracy of smFISH is shown to be around 80% [29]. Another smFISH approach using multiple colour barcoding enables quantification of multiple mRNA targets at once, and even discriminates single nucleotide variants and is also able to distinguish newly synthesized RNA from mRNA [36–38].

Recently, a variant of smFISH using branched DNA technology has been used to obtain image based transcriptomics [39, 40]. Besides measuring the number of mRNA transcripts in single cells, this image-based approach enables measuring variables like total protein content, total mRNA content, cell size, the number of neighbouring cells, and numerous additional variables.

The accuracy of mRNA detection is comparable to the accuracy reported for normal smFISH [29, 40]. Another use of smFISH is multiplexed error-robust fluorescence in situ hybridisation (merFISH). merFISH enables detection of the copy numbers and spatial localization of up to 1000 different mRNA species in single cells. The detection of the different mRNA species is achieved by sixteen sequential hybridizations with encoding schemes targeting different parts of each mRNA in the different hybridization rounds [41].

The main advantage of single-cell in situ transcription detection is the additional information about the spatial localization of the detected mRNA molecules. Besides microscopy-based techniques, various other methods to quantify single cell expression patterns have been established (see Kalisky et al. [42] for a review). These techniques have previously been applied to assay population-wide characteristics. For instance, it has been shown that profiling of gene expression in single cells can be performed with quantitative PCR (qPCR) [43–45]. However, in a single cell setting, the main hurdle is the use of small quantities of target molecules. The conversion of mRNA to cDNA during RNA-seq methods appears to be a limiting step [46]. It has been estimated that the mRNA detection efficiency is only around 10% [47]. Interestingly, Wu et al. showed that the obtained expression levels using single-cell RNA-seq and single-cell qPCR correlated significantly [45]. The main advantage, however, of single-cell RNA-seq is the detection of expression levels of (in theory) the full transcriptome. Besides all mRNA, the detection of non-coding RNA and different splice variants might also be possible [42]. Recently, Lee et al. [46, 48] showed how the benefits from both approaches, i.e. spatial localization and detecting a multitude of target genes, can be combined in a technique referred to as fluorescent in situ RNA sequencing (FISSEQ).

The current trend of gene expression variability detection is towards a combined assessment of multiple mRNA targets or quantification of mRNA in combination with measurements of the analyzed cell state. Such integrated approaches provide a more complete view on gene expression and enable determination of the cellular parameters that drive cell-to-cell variability in gene expression. One such study showed that the expression state of genes can be predicted by a collection of phenotypic variables [40]. Of course, such relations can only show correlations and do not provide the causal link between measured variables. Modeling approaches can be helpful to infer causalities and real-time behavior of the involved parameters. However, to fully unravel how various components contribute to cell-to-cell variability in gene expression mechanistically, it will be essential to measure the involved processes in real-time.

1.8 Thesis outline

Gene expression levels of isogenic cells can show large cell-to-cell variability thereby allowing cells to display defined cellular phenotypes. The mechanisms that contribute to distinct gene

expression states in isogenic cells are of fundamental importance to understand cellular functioning and behavior in response to environmental changes. Different hierarchical levels of gene regulation can be identified that all contribute to the total variability in gene expression. In this thesis, we study how different regulatory levels of gene expression each contribute to variability in single cells. In this thesis, gene expression is measured in single cells using techniques to quantify transcripts at the single molecule resolution level. Moreover, theoretical and computational approaches are used to describe and understand gene expression cell-cell variability.

In chapter 2 we illustrate that our approach to determine mRNA concentrations in single cells in combination with volume measurements provides a more robust measure of gene expression cell-to-cell variability compared to measuring absolute transcript levels. By combining single-cell mRNA expression measurements with cell volume measurements, we quantify the mRNA concentration of a reporter gene embedded in a known genomic context in single cells. We observe that in a population of growing cells the mRNA copy number per cell scales proportionally with cell volume. As a result, the mRNA concentration remains constant over different cell volumes, indicating that the mRNA concentration is in homeostasis during cell growth. This observed mRNA concentration homeostasis indicates that although the promoter of the analyzed reporter gene is considered to drive constitutive gene expression, cell volume (or cell cycle age) impacts either gene expression or mRNA degradation constants of the studied gene. In conclusion, our study shows that biologically relevant variation in gene expression is best captured by concentration variability.

In chapter 3 we describe a stochastic simulation tool to study gene expression in a growing cell population. We show that a previously derived theoretical framework for balanced growing cell populations is able to describe single cell growth. Our experimental data of single bacterial cells in a cell population in balanced growth confirm our theoretical findings. We use this theory to construct a model that simultaneously simulates cell growth and gene expression. The simulations are tested on protein expression data in the same bacterial cell populations. The model enables simulations and analyses of elaborate gene expression networks of growing cell populations. This chapter lays the groundwork for developing and testing hypotheses on the impact of cell growth on gene expression networks. We provide a tool to study remaining questions from the previous chapter, i.e. how cells cope with DNA concentration dilution due to cell growth and DNA replication, and how cells maintain concentration homeostasis.

Besides homeostasis of mRNA concentrations, we observed differences in transcription variability between the cell lines studied in chapter 2. These differences in transcription variability can be explained by differences in the local genomic context of the studied genes. In chapter 4 we determine the effect of altering the chromatin state on transcription repression dynamics in single cells in real-time. We use a mammalian cell line containing a reporter gene cassette that expresses MS2-tagged mRNA molecules. Our data shows that upon removing the transcriptional activator from the reporter gene transcription initiation continues transiently whereas

the transcriptional activator is instantaneously lost from the reporter gene cassette. We show a biphasic transcription repression response representing the actual decrease in the transcripts numbers at the gene array and a preceding response time. Binding of the epigenetic reader protein methyl-CpG-binding protein 2 (MeCP2) accelerates the transition rate of transcription repression significantly. However, the diffusion rate at which the transcripts diffuse from the reporter gene cassette in the nucleus is unaffected. We hypothesize that MeCP2 acts as a facilitator to provide transcription repression.

In chapter 5 we study the impact of environmental changes on gene expression. Specifically, we study the role of UV-exposure-induced DNA damage on transcription stalling and transcription recovery after nucleotide excision repair. We studied the effect of UV-induced damage on transcription stalling comparing the effect of exposure to UV with transcriptional inhibitor treatment. We use single-molecule mRNA FISH to measure on transcription at the single gene level upon inducing UV-induced damage. We show that gene size and UV-dosage impact gene expression recovery after UV-damage. Smaller genes are less prone to acquire DNA damage after UV-exposure. Measuring the expression levels in time, combined with a computational DNA damage gene expression model, allows us to estimate the transcription recovery after UV damage to infer the transcription recovery rate after UV exposure. Our results suggest that dependent on the UV-dose different DNA-damage detection and repair mechanism are active.

In chapter 6 we study the effect of variability in gene expression in human breast cancer cells. Approximately 70% of diagnosed breast cancers are estrogen receptor positive. Estrogen receptor positive breast cancer can be treated with aromatase inhibitors which prevent endocrine signaling and with it proliferation. We demonstrate using smRNA FISH that in breast cancer cells treated with aromatase inhibitors, a subpopulation of cells overexpresses aromatase. The smRNA FISH approach allows us to determine the presence of single cells with an irregular transcription state. Our study suggests that treatment resistance may be initiated by a single or a few cells circumventing the treatment effects.