Regulation and Cell-to-Cell Variability of Estrogen-Dependent Transcription

Dissertation

zur Erlangung des Grades

"Doktor der Naturwissenschaften"

am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

Christoph Alexander Fritzsch

geboren am 12.03.1986 in Jena

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Mainz, 2017

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 03.12.2018

Abstract

Transcription is the major control point for the production of cellular components, and as such, it governs fate and function of cells. Surprisingly, this critical step in information transmission is rather unreliable: The low copy-number of molecules controlling transcription, leads to stochastic effects in the production of RNAs. Discontinuities in transcriptional permissiveness of gene promoters, witnessed as “transcriptional bursts”, amplify this effect and lead to variability in gene expression between cells. Furthermore, the cellular state influences the capacity of transcription in a cell-specific manner, thereby affecting transcriptional output on long timescales. Variegated expression can be beneficial in development and cellular decision-making. However, it is also exploited by cancerous tissue, as intra-tumor diversity lowers therapeutic effectiveness. It is essential to study stochastic transcriptional regulation at the single-cell level, to comprehend sources and consequences of cellular heterogeneity in health and disease.

This thesis examines how cells control transcriptional bursts to adapt gene expression output to environmental signals and how the cellular state influences this process. An estrogen-sensitive locus served as a model system as it allows for tight experimental control of expression through varying estrogen levels. Fluorescent labeling of nascent transcripts in combination with time-resolved microscopy in living cells enabled quantitative measurements of single-cell transcriptional dynamics. Transcription occurred in stochastically timed bursts with a strong cell-to-cell variability in long-term transcriptional output. Stochastic mathematical models of promoter progression and transcription were fitted to the acquired datasets to discriminate alternative hypotheses of promoter regulation. This revealed that estrogen adjusts the frequency of transcriptional bursts by controlling the transition to a transcriptionally permissive promoter. The cellular state, however, alters long-term transcriptional output through initiation and elongation kinetics. This effect is mediated through a diffusible factor, as two alleles within the same cell were similarly affected and recently divided daughter cells correlated in time-averaged transcription.

To infer whether the chromatin environment influences burst characteristics and transcriptional noise, perturbations using inhibitors of chromatin modifying enzymes were performed. Interestingly, the inhibition of histone deacetylation reduced noise in gene expression, highlighting that chromatin permits noise regulation at the level of nascent transcription. In conclusion, this thesis provides a quantitative description of estrogen-dependent transcription, which incorporates transcriptional bursting, the influence of cellular state, and gene-specific tuning through chromatin.

Zusammenfassung

Der Prozess der Transkription bestimmt maßgeblich das Schicksal und die Funktionsfähigkeit einer Zelle, indem er die Produktion zellulärer Bestandteile steuert. Überraschenderweise ist dieser Schritt der Informationsübertragung nicht sehr zuverlässig: Da die an der Transkription beteiligten Moleküle nur in geringer Anzahl in der Zelle vorliegen, entstehen stochastische Effekte bei der Herstellung von RNAs. Diese werden verstärkt durch zeitliche Fluktuationen in der Transkriptionsfähigkeit eines Gens, welche als transkriptionelle „Bursts“ zu beobachten sind, wodurch sich Genexpressionsmuster von Zelle zu Zelle unterscheiden. Außerdem wirkt sich der interne Zustand einer Zelle auf die zellspezifische Transkriptionsaktivität aus. Die daraus resultierende Expressionsvariabilität kann während der Gewebeentwicklung und der zellulären Entscheidungsfindung vorteilhaft sein. Allerdings führt die Heterogenität von Krebszellen im Tumorgewebe auch zu einer Beeinträchtigung des therapeutischen Erfolgs. Die Untersuchung stochastischer Effekte ist daher wichtig, um die Entstehung zellulärer Heterogenität und deren Bedeutung für gesundes sowie krankes Gewebe zu verstehen.

In dieser Arbeit wurde zum einen untersucht, wie transkriptionelle Bursts reguliert werden, um Höhe und Variabilität von Geneexpression an Umweltsignale anzupassen und zum anderen, wie dieser Prozess durch den inneren Zustand der Zelle beeinflusst wird. Dabei wurde ein östrogenabhängiges Gen als Modellsystem gewählt, um eine gezielte Steuerung der Expression durch Änderung der Hormonmenge zu ermöglichen. Naszierende RNAs wurden durch Fluoreszenzmarkierung im Mikroskop sichtbar gemacht und über verschiedene Zeitpunkte in lebenden Zellen verfolgt. So konnte die dynamische Transkriptionsaktivität einzelner Zellen und die daraus resultierende Variabilität quantifiziert werden. Transkription trat in stochastischen Bursts auf, wobei sich die mittlere Transkriptionsaktivität von Zelle zu Zelle unterschied. Stochastische Modelle, welche die zeitliche Entwicklung des Promotorzustandes und des Transkriptionsprozesses beinhalteten, wurden an die aufgenommenen Daten angepasst und erlaubten es, unterschiedliche Hypothesen der Promotorregulation zu testen. Hierbei ergab sich, dass Östrogen die Frequenz transkriptioneller Bursts ändert, indem es den Übergang in einen aktiven Promoterzustand reguliert. Der Zellzustand hingegen beeinflusst die Genexpression durch Änderung der Initiations- und Elongationsraten. Dafür ist ein diffundierender Faktor verantwortlich, da sowohl zwei Allele innerhalb derselben Zelle, als auch Tochterzellen nach der Zellteilung in der zeitgemittelten Transkription korrelierten.

Um herauszufinden, ob Chromatin die Eigenschaften von Bursts und die resultierende Expressionsvariabilität beeinflusst, wurden chromatinabhängige Prozesse inhibiert. Interessanterweise reduzierte eine Inhibition der Histondeacetylierung die Variabilität der Genexpression. Chromatin ermöglicht also eine Regulation der Heterogenität bereits während der Produktion von RNAs. Schlussendlich bietet diese Arbeit eine quantitative Beschreibung östrogenabhängiger Transkription, die sowohl transkriptionelle Bursts und den Einfluss des zellulären Zustands, als auch genspezifische Regulation auf Chromatinebene beinhaltet.

Table of Contents

1 Introduction

1.1 Heterogeneity in biology

"This preservation of favourable variations and the destruction of injurious variations, I call Natural Selection, or the Survival of the Fittest. Variations neither useful nor injurious would not be affected by natural selection and would be left a fluctuating element."

— Charles Darwin, *Origin of Species* (5th edition, 1869)

Heterogeneity (from Greek: *heteros* "different", *genos* "kind") is a hallmark of biology and describes that observable individual entities differ from each other within a population. Consequently, deviations from the mean are the norm rather than the exception, and indeed, diversity exists in biology at various length- and time-scales (Figure 1).

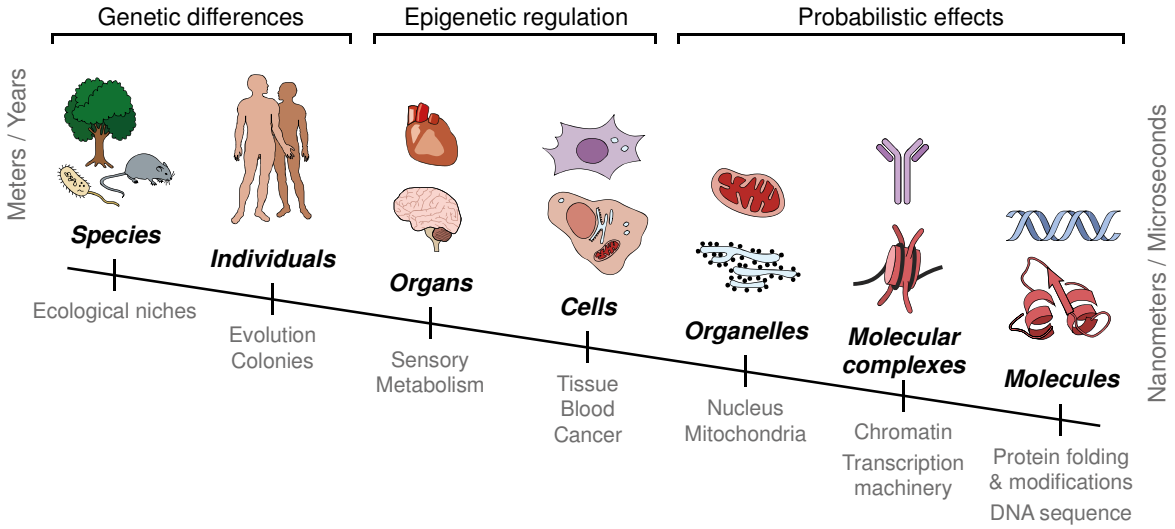

Figure 1: Heterogeneity is a hallmark of biology across scales. Variability exists in numerous examples, from distinct species that evolved in millions of years down to molecules that take on various shapes in a fraction of a second. Genetic and epigenetic differences cause differences at medium-to-large scales, while probabilistic events are predominant at the molecular level.

An enormous amount of distinct structures, behaviors, and interactions has evolved that allowed adaptation of species to ecological niches and the formation of complex ecosystems. *Within* a species, individuals differ in sex, appearance, development, behavior, and more. While these differences are often due to slight alterations in the genetic content, cells inside a multicellular organism are genetically homogenous—and yet they take on a multitude of shapes and functions. The ability of a single fertilized oocyte to differentiate into specialized cell types facilitates formation of tissues and organs with specific functions. Recursively, variability is even apparent *within* an individual cell, through organelles that provide compartments for distinct biochemical pathways, through differential assembly of molecular complexes that allow a enzymes to specifically act on multiple targets, or even at the level of individual molecules that can take on various shapes, e.g. through isoforms or folding.

Biological variability has several benefits: First and foremost, it allows natural selection to occur. Only if individuals of a population differ, and when these differences are heritable, selection of the best-adapted phenotypes is possible. As such, heterogeneity is *the* major driver of evolution. A second benefit of variation is specialization and division of labor. On the level of individuals, a fascinating example are colonies of eusocial insects, where individual animals take on specialized behaviors, like reproductive queens and sterile workers. Moreover, the development of organ systems and the use of organelles within cells highlights that separation of tasks to specialized units is also beneficial within organisms. At the level of individual cells, heterogeneous responses to environmental cues can also aid in biological decision making, for example during differentiation (Chang et al. 2008).

Conversely, too much variation is detrimental and hinders biological function. For instance, two distinct species cannot produce fertile offspring, and fitness generally decreases with variation (Wang & Zhang 2011). Variability also plays a role in disease: Heterogeneity in bacterial populations allows for antibiotic resistance (Balaban 2004). Abnormal gene expression within an individual cell can cause abnormal growth and the formation of tumors. Cancer cells even exploit natural variation to overcome barriers in proliferation as the tumor evolves (Brock et al. 2009, McGranahan & Swanton 2017) and escapes therapy.

What are the mechanisms by which such widespread heterogeneity arises and how can it be controlled? At large scales, differences can be attributed to changes in genomic information. Species have evolved different genomes over millions of years, and individuals within a species also differ in DNA sequence by single nucleotide polymorphisms and spontaneous mutations. As genomic differences are inheritable, they are also the driver behind evolution. However, within each individual, all cells essentially share the same genomic information (with exceptions in the immune system and rapidly mutating cancer cells). Hence, differences between cell types cannot be attributed to genetic heterogeneity and are non-genetic in nature.

1.2 Non-genetic heterogeneity

Within an individual, non-genetic regulatory mechanisms dominate. They control which parts of the genome are being expressed in a cell. It is this specific gene expression pattern that distinguishes cell types throughout development and within different organs. Cellular identity is stably maintained, but in addition, cells can dynamically adjust expression to internal (e.g. cell cycle) and external (e.g. growth factors, hormones) signals (Perkins & Swain 2009). Regulation of gene transcription is hence the most important regulator of cell behavior. This regulation occurs through the set of transcription factors that a cell contains in combination with epigenetic alterations that allow stable maintenance of transcriptional programs. These mechanisms are described in detail in section 1.3.

A further contributor to non-genetic heterogeneity is randomness, an inherent property of every biochemical process. It mainly acts at the molecular scale and arises through stochasticity in molecular interactions. While this effect is averaged in processes with large

molecule counts, e.g. metabolism, it has a major impact where the number of participating molecules is low; for example, when transcription factors bind to a single specific site within the genome in order to regulate expression of a target gene. Consequently, even isogenic cells of identical cell type grown under identical conditions can differ remarkably in their levels of specific RNAs and proteins. The effect of stochasticity on gene expression is further described in section 1.4. Probabilistic effects also occur on larger scales, for example, during cell division, when constituents and organelles are randomly partitioned between daughter cells. How cells coordinate development and reliably maintain cell identity in the light of uncertainty of molecular interactions is fascinating. It is one aim of this thesis to understand non-genetic heterogeneity and its effects on gene regulation.

1.3 Transcriptional regulation and chromatin biology

The DNA within each cell of a multicellular organism contains the same genome, with all information that is necessary to build the entirety of RNA and protein molecules of a body. Nevertheless, the hundreds of cell types within our body differ dramatically in appearance, content and function. The basis of this diversity is how the genomic information is accessed and hence, which subset of the genome is used to produce RNAs and proteins. The general flow of information is from DNA to RNA then to protein, with multiple intermediate processing steps required to produce a functional molecule. While all of these steps are tightly regulated, the most critical regulatory event is the production of RNA from the genomic template in the process of transcription. Transcriptional control itself is a multilayered process that consists of cell-specific transcription factors, which regulate individual genes based on sequence-specific DNA binding, in conjunction with epigenetic mechanisms, which regulate DNA accessibility through meta-stable chromatin structures. This thesis aims to understand the regulation of gene expression and associated stochasticity with contributions of the chromatin environment around a native gene locus.

1.3.1 Regulation by transcription factors and *cis*-regulatory elements

General transcription factors are needed for transcriptional activation of most genes. They assemble at the core promoter upstream of the transcriptional start site (TSS) of a gene, separate the two anti-parallel strands of DNA, then recruit, position and phosphorylate the C-terminal domain (CTD) of RNA Polymerase II such that transcription commences (Lee & Young 2000) (Figure 2). Basal gene activity, as mediated by general transcription factors, however, is quite low. Sequence-specific transcription factors (TFs) can increase transcriptional output by several orders of magnitude through binding to DNA at distal regulatory elements, called enhancers (Shlyueva et al. 2014). These are short (<1000 bp) sequences containing multiple binding sites for TFs. As such, enhancers integrate the information of several TFs with activating and repressing cofactors in a combinatorial fashion to regulate gene expression in a spatially and temporally confined way. Furthermore, a single gene can be regulated by multiple enhancers and a single enhancer can regulate multiple genes, thereby, increasing the regulatory complexity. Enhancers can influence expression of genes over large distances (several kilobases) and act independent of their orientation. Long-range activation requires formation of chromatin loops to bring enhanc-

ers and promoters in close proximity, often mediated by cohesin and mediator complexes. While it is unclear how enhancers select their specific target promoters from a large pool of possible sequences, higher-order chromatin structure restricts the repertoire of possible interactions. Insulator sequences separate the genome into megabase-sized topologically associated domains (TADs), within which interactions occur, but between which only few interactions take place (Dixon et al. 2012).

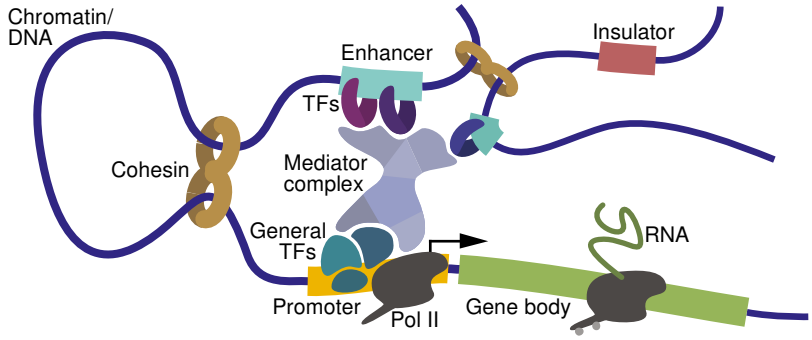

Figure 2: Gene regulation by sequence-specific transcription factors and *cis*-regulatory elements. Transcription factors (TFs) recognize binding sites in enhancer regions and recruit coactivators and the mediator complex. Chromatin looping via cohesin mediates contact with the core promoter, where general TFs are in contact with the mediator and aid in melting of DNA and recruitment of RNA polymerase II (Pol II) to initiate transcription.

The human genome encodes for about 1800 different transcription factors (Vaquerizas et al. 2009) and the specific set of transcription factors that are expressed in a cell determines its transcriptional program and the capacity of each cell to respond to external and internal signals. All of these transcription factors regulate each other's expression within a gene regulatory network with complex feedback systems (Babu et al. 2004). Such networks on their own can lead to intricate kinetic behavior, like oscillations (Doherty & Kay 2010, Elowitz & Leibler 2000), but can also stably maintain a cells expression state despite changes in the cellular microenvironment, thereby establishing memory in cellular decision making (Burrill & Silver 2010). In addition to regulatory networks, gene expression is controlled at the level of chromatin accessibility.

1.3.2 Chromatin control of gene expression

The approximately two meters of genomic DNA contained within each nucleus of a human cell are highly compacted and organized by histone and non-histone proteins into a dynamic macromolecular structure called chromatin (Figure 3). The basic unit of chromatin is the nucleosome core particle, which wraps 146 bp of DNA around an octamer of histone proteins, composed of two histone H2A-H2B dimers and an H3-H4 tetramer (Luger et al. 1997). The mobile linker histone H1 then promotes further compaction of the chromatin into a higher-order structure of 30 nm fibers (Robinson & Rhodes 2006). The positively charged amino acid residues within the small basic histone proteins form a tight electrostatic interaction with the phosphate backbone of DNA. This masks the underlying DNA sequence and makes it inaccessible to cellular regulators; for example, the binding of transcription factors. In general, chromatin establishes a highly repressive environment, where multiple regulatory barriers must be overcome for transcription to occur.

In addition to structural functions, histones act as information hubs that integrate multiple signals. Attaching or removing posttranslational modifications of the amino-terminal tails and within the globular core of histones, remodeling of nucleosomes, or the incorporation of histone variants allow for formation of meta-stable chromatin “states” that relay information to other cellular machinery. An extreme example is the formation of transcriptionally inactive and densely organized heterochromatin, as compared to the more open and transcriptionally permissive euchromatin.

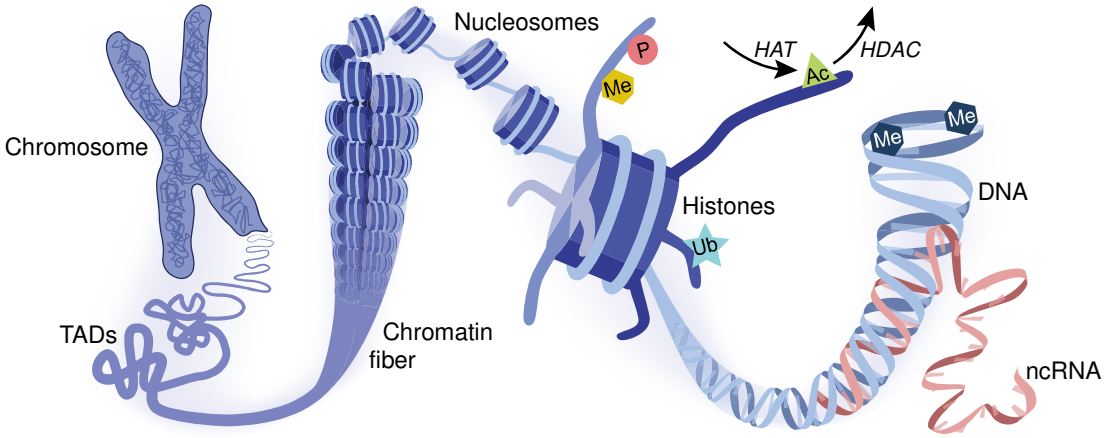

Figure 3: Chromatin organization. DNA is wrapped around nucleosomes, which are arranged like “beads on a string” and form higher order structures, e.g. topologically associated domains (TADs). Histones carry a variety of post-translational modifications with impact on DNA accessibility and binding of regulatory proteins. DNA methylation and non-coding RNAs are additionally involved in regulating gene expression.

Posttranslational modifications of amino acid residues are widespread in all histones and affect chromatin in two ways: either they directly modify the overall structure of chromatin, or they affect the binding of effector molecules. About 130 different histone marks have been identified so far (Tan et al. 2011). They include acetylation, methylation, phosphorylation, deimination, sumoylation, ubiquitination, crotonylation, and ADP ribosylation of amino acids (Kouzarides 2007), each of which occurs at multiple histone residues. A single nucleosome can carry various modifications, with their combination defining a “histone code” that determines DNA accessibility and the ability to recruit specific chromatin “readers” (Jenuwein & Allis 2001, Strahl & Allis 2000).

Histone acetylation is generally associated with an open chromatin structure that is permissible for transcription. This is presumably because the transfer of an acetyl group from acetyl coenzyme-A to the ε-amino group of lysine (K) residues, neutralizes their positive charge and weakens the interaction between histone and DNA. Histone acetyl transferases (HATs) and histone deacetylases (HDACs) are the enzymes that catalyze the transfer and removal of acetyl groups on histones, respectively. Both enzymes also have non-histone targets, and can also act on longer side chains, like propionylation and butyrylation (Chen et al. 2007, Kebede et al. 2015). HDACs are classified into four distinct groups: Class I, II, and IV are zinc-dependent and can be inhibited by hydroxamic acids like Trichostatin A (TSA) and the carboxylic acid butyrate. In contrast, class III HDACs, also called sirtuins, use nicotinamide adenine dinucleotide (NAD+) as a cofactor. In addition to

altering direct DNA interaction, acetyl residues are recognized by chromatin readers via bromodomains or tandem PHD domains (Yun et al. 2011).

Methylation of histone tails has divergent functional roles, depending on the specific modified residue and the methylation state (mono-, di-, or tri-methylation on lysines, and mono- or di-methylation on arginines). While tri-methylation of H3K9, H3K27, and H4K20 are hallmarks of heterochromatin, the methylation of H3K4, H3K36, and H3K79 are correlated with transcriptional activity (Greer & Shi 2012). Methylation marks are set through the action of histone methyl transferases (HMTs) and removed by lysine demethylases (LSDs). Another modification on histones is phosphorylation, which changes the charge of histone proteins by adding a negatively charged moiety to serine and threonine side chains, and signals by recruitment of downstream effectors. For example, H3S10 phosphorylation is involved in chromatin compaction during mitosis and transcriptional activation (Nowak & Corces 2004), while H2AXS139 phosphorylation is involved in DNA repair (Rogakou et al. 1998).

In order to overcome the repressive barrier of chromatin and permit binding of transcription factors and polymerases, nucleosomes can be moved and detached from DNA. This process is mediated by the activity of chromatin remodeling complexes in an ATP-dependent manner (Wang et al. 2007). Chromatin remodeling is also involved in the replacement of canonical histones by histone variants. Slight differences in amino acid composition in these variants further increase the signaling capacity of chromatin. H2AX for instance, is incorporated after a DNA double strand break, while H3.3 marks actively transcribed genes, and CENP-A is an H3 variant that marks centromeres (Biterge & Schneider 2014).

In addition to histones, two more constituents of chromatin are involved in regulating gene expression: DNA itself and non-coding RNAs. DNA can be modified at cytosine bases by attaching a methyl group at the 5’ position of the pyrimidine ring. 5-methylcytosine occurs in the context of CpG dinucleotides, which are enriched at gene promoters and enhancers, with methylation correlating with gene silencing. DNA methylation regulates genes during genomic imprinting (Li et al. 1993), retro-element silencing (Walsh et al. 1998), and differentiation (Bock et al. 2012). Further oxidation of the methyl group leads to 5-hydoxymethyl-, 5-formyl-, and 5-carboxycytosine. These are intermediates in oxidative DNA demethylation through eventual base-excision repair. In addition, oxidative intermediates may also carry out regulatory functions (Dahl et al. 2011).

Non-coding RNAs (ncRNAs) influence gene expression at various stages. On the one hand, they fulfil critical structural and enzymatic roles during splicing (small nuclear RNAs) and translation (transfer RNAs, ribosomal RNAs). On the other hand, they also regulate mRNA levels. Short ncRNAs, like micro RNAs, small interfering RNAs, or piwi-interacting RNAs are involved in post-transcriptional gene silencing (Patil et al. 2014), while long ncRNAs function by recruiting chromatin modifiers to regulate transcription directly, for example during X-inactivation (Chow & Heard 2009).

This multitude of epigenetic regulatory mechanisms allows precise control of gene expression during differentiation and in response to environmental cues. The relative stability of the chromatin environment establishes maintenance of developmental decisions and the preservation of cellular identity. In contrast, the chromatin state around regulated promoters can also be quite dynamic, with remodeling, addition, and removal of histone and DNA modifications occurring within minutes upon transcriptional activation (Kangaspeska et al. 2008, Métivier et al. 2003).

1.4 Stochastic gene expression and transcriptional bursting

Many cellular functions critically rely on the interaction of molecules that are present at a few copies per cell. In consequence, processes dependent upon low abundance factors have an inherent uncertainty, or noise, in timing or outcome of reactions. This gives rise to stochastic fluctuations over time. One remarkable example of stochastic behavior is the control of gene expression, where many stochastic molecular interactions play a role in the process of transcription: Transcription factors have to find specific binding sites within the genome, and then recruit further cofactors and polymerases, while all of these factors diffuse randomly through the nucleoplasm. Similarly, each interaction of a mature mRNA with the ribosome or with the degradation machinery is a product of random encounters of molecular complexes. Hence, many steps on the way from gene to protein introduce uncertainty and provide a stochastic outcome.

The amount of randomness scales with system size: At low abundance, even the random birth or death of a single molecule severely influences its concentration, while these effects are negligible at high abundance. This finite-number effect leads to scaling of noise (measured by the squared coefficient of variation (CV^2) = variance/mean^2) with the inverse of the number of molecules (Kaern et al. 2005, Paulsson 2004, Raj et al. 2006). A single transcription factor binding site within a promoter of a target gene or low absolute numbers of an mRNA are examples where the finite-number effect impinges upon outcome. Low mRNA numbers, in combination with high translation rates, produce strong mRNA fluctuations, which propagate through to protein levels, leading to high protein noise. In contrast, a regime with the same protein abundance, but achieved with low translation rates from high mRNA numbers shows less fluctuation in mRNA levels and less protein noise. Hence, cells can control protein noise through tuning of transcription and translation rates (Thattai & van Oudenaarden 2001). In addition, stochasticity already arises at the level of RNA production and mathematical models were developed to quantify the contribution of transcription on expression noise.

1.4.1 Models of stochastic gene expression

Mathematical models of gene expression are helpful to describe the dependence of noise on kinetic parameters of transcription. As such, they provide an assessment of whether experimental observations of noise are in agreement with a proposed model and allow discrimination of alternatives. For example, a low noise level is expected for the simplest model of transcription, in which the promoter of a gene is continuously active and always

able to initiate polymerases—when chromatin is always open or the transcription factor is constantly bound. In this case, the RNA concentration can be described by assuming that production and decay follow single rate-limiting steps (Figure 4A, scheme). For such a birth-death process, the steady-state RNA distribution follows a Poisson distribution (Sanchez et al. 2013), which is characterized by low noise, with the standard deviation being equal to the mean (hence, CV^2 = 1/mean). Such a one-state promoter model was for example able to describe the expression of housekeeping genes in yeast (Zenklusen et al. 2008).

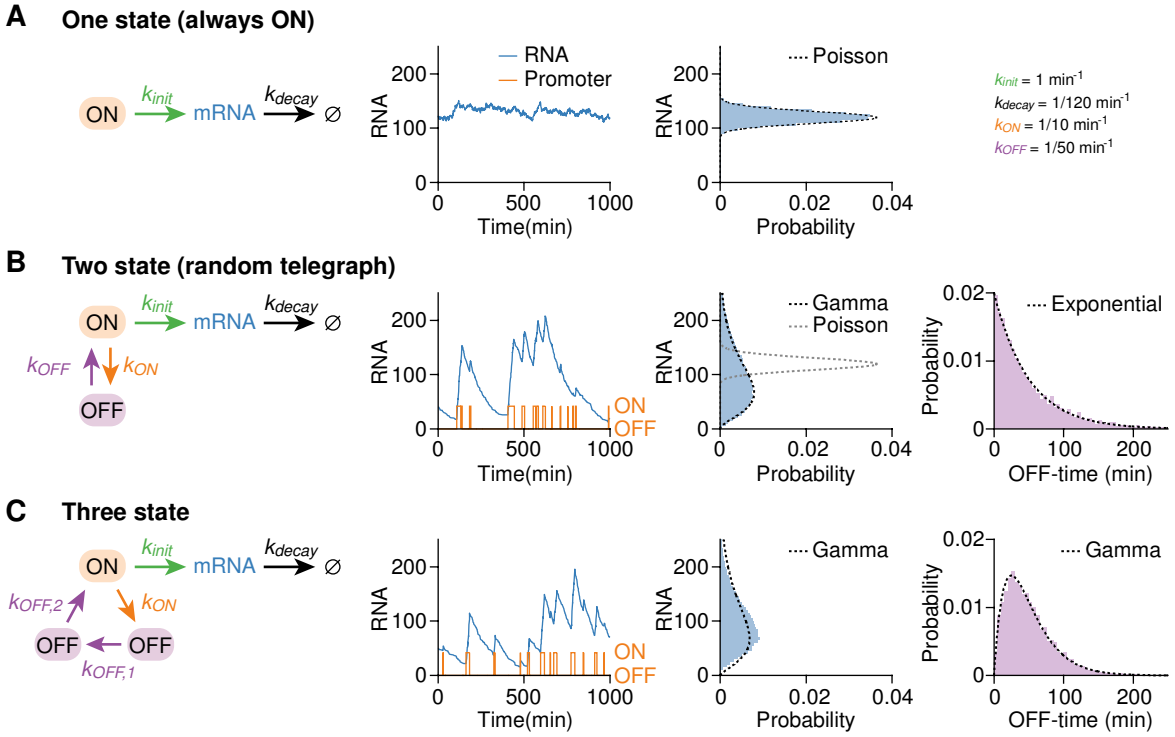

Figure 4: Qualitative differences in the distribution of RNA numbers and OFF-times for three models of stochastic gene expression. Model topologies are shown for different models of gene expression (left). Transcripts are produced with a rate k_{init} and decay with a rate k_{decay}. Switching between promoter states occurs with rates k_{ON} and k_{OFF}. Stochastic simulations are shown for each model (middle-left) along with the simulated RNA (middle-right) and OFF-time distributions (right). Dashed lines indicate analytical distributions. **(A)** For a promoter that is always active, transcripts show little fluctuations around a mean level, leading to a Poisson distribution in RNA copy number. **(B)** Strong transcript fluctuations are observed for a two-state model leading to more variability in transcript numbers (approximated by a gamma distribution). OFF-times are exponentially distributed because they depend on a single step. **(C)** The presence of two OFF-states in the three-state model leads to a slightly narrower transcript distribution (the same Gamma distribution as in panel B is shown) and a peak in the OFF-time distribution.

In contrast, much larger noise levels (standard deviation >> mean) were measured for most eukaryotic genes (Larson et al. 2013, Raj et al. 2006), indicating another layer of stochasticity. It was suggested that the promoter is not constantly active, but rather switches between transcriptionally active and inactive states with DNA/chromatin structure or protein occupancy defining transcriptional activity of such a state (see 1.4.2). The discontinuity in RNA production that arises from promoter switching is termed transcriptional "bursting". The simplest realization is a two-state or "random telegraph" model (Paulsson 2005 and Figure 4B) with one active and one inactive promoter state. It leads to much

higher cell-to-cell variability, with transcript levels following a wider distribution, which, in the case for infrequent and short bursts, is represented by a gamma distribution (Sanchez et al. 2013). Such distributions were indeed observed for mammalian genes (Raj et al. 2006). Furthermore, direct evidence for bursting was delivered by dynamic studies of nascent transcription in living cells, which showed that transcripts were produced in short time intervals, interspersed by long silent periods (Chubb et al. 2006, Golding et al. 2005).

A two-state promoter model introduces further points of control for transcriptional output as compared to the one-state model. In addition to the transcription initiation rate, the rates at which a gene switches between active and inactive periods can be tuned to achieve different mean transcript levels. Transcriptional bursting therefore enables two major ways to tune gene output: modulation of burst size and modulation of burst frequency. Burst size modulation affects the number of transcripts that are being produced per burst by either changing the burst duration (ON-time) or the initiation rate from the ON-state. Frequency modulation is realized when the time in between bursts, i.e. the OFF-time, is altered, giving rise to more or less bursts per time interval without affecting the number of transcripts that are produced per burst. At the same expression level, a gene with long OFF-times and high burst size shows a higher noise level than a gene with fast promoter switching and lower burst size (more similar to always ON). Hence, promoter kinetics influence noise at the level of nascent RNA.

The multitude of protein and RNA factors that are involved in gene activation and polymerase firing (Figure 2), in combination with the dynamics of chromatin (Figure 3) suggest that the molecular mechanism of promoter switching between inactive and active states might be more complex than a simple one-step process. Indeed, many mammalian genes show a refractory period for gene activation that is indicative of multiple steps in the inactive phase (Harper et al. 2011, Suter et al. 2011, Zoller et al. 2015). Such multi-state models show a non-zero peak in the distribution of waiting times between bursts (Figure 4C). Hence, the switching between active and inactive phases is more regular and the cell-to-cell variability in transcript numbers is reduced. By fitting a set of models of promoter progression to experimental data, this thesis aims to reveal the structure and kinetic parameters of promoter switching at various induction levels. This enables inference of regulatory principles for stimulus dependence (burst frequency vs. burst size modulation) and predictions on noise scaling.

1.4.2 Mechanisms of transcriptional bursting

While it is now accepted that transcriptional discontinuity is widespread and occurs in bacteria to human, it is less clear what the molecular determinants of activity and inactivity of a promoter are. It is likely that mechanism differ between genes and organisms (Lenstra et al. 2016, Nicolas et al. 2017). In bacteria, for example, the buildup of supercoiling during transcription can lead to pauses in between gene activity, during which supercoiling is removed (Chong et al. 2014). In eukaryotes, chromatin sets a barrier for transcription and allows for gene-specific regulation in combination with promoter architecture. Several mechanisms are possible that result in transcriptional bursting: (I) Nucleosome positioning

can regulate the accessibility of transcription factor binding sites with remodeling being responsible for switching between promoter states. (II) The transcription factor itself can be involved during a burst when initiation only occurs as long as the transcription factor is bound or when the factor shows pulses in nuclear localization (Cai et al. 2008, Hao & O'Shea 2011). (III) Furthermore, re-initiation of polymerases from an assembled pre-initiation complex can result in a burst in RNA production. (IV) Finally, elongation arrest and interplay between polymerases during elongation can lead to several polymerases traveling along the gene in close proximity with synchronous transcript release (Fujita et al. 2016). Probably, these mechanisms do not act in isolation but rather interact and co-occur at the same gene, giving rise to complex bursting behaviors.

1.4.3 Intrinsic and extrinsic contributions to gene expression noise

The stochasticity in promoter switching and polymerase initiation gives rise to considerable noise in gene expression. However, this is only a part of the total observed noise. Total noise can be separated into contributions from fluctuations that are inherent to the process of interest, termed intrinsic noise, and contributions form external factors, termed extrinsic noise. In the case of gene expression, intrinsic noise is generated by the stochasticity inherent to the reactions, for example the birth or death of individual molecules, while extrinsic noise defines fluctuations in the associated rate constants (Kaern et al. 2005). For instance, the number of polymerases or ribosomes, which determine the rate of transcription or translation, respectively, could be subject to fluctuations. Which processes are termed intrinsic or extrinsic also depends on the context and has to be defined individually (Figure 5). For instance, "gene-intrinsic" noise describes noise associated with promoter switching and the birth and death of individual RNAs (or proteins) from a single gene. "Pathway-intrinsic" noise incorporates fluctuations in upstream signaling molecules with effects on multiple genes. "Cell-intrinsic" noise additionally includes the specific state of the cell, for example cell cycle stage, energy content, or the number of polymerases and ribosomes, affecting all genes.

In this thesis, the term intrinsic noise is used for noise that arises solely at the level of stochastic promoter switching and the stochastic initiation of polymerases. Extrinsic noise describes all processes that influence kinetic rates of gene expression up- or downstream of initiation. For example, fluctuations in the level of transcription factors that determine promoter dynamics, or factors that drive transcript elongation. Usually, extrinsic noise acts on longer timescales because cellular state is thought to fluctuate slower than individual bursts of RNA production. Hence, time averaging can provide an estimate of extrinsic noise. However, the gold-standard for estimation of intrinsic and extrinsic contributions is the observation of two identical reporter genes within the same cell (Elowitz et al. 2002). Extrinsic noise is then apparent as the correlation in gene products from these two reporters (compare Figure 5B and C).

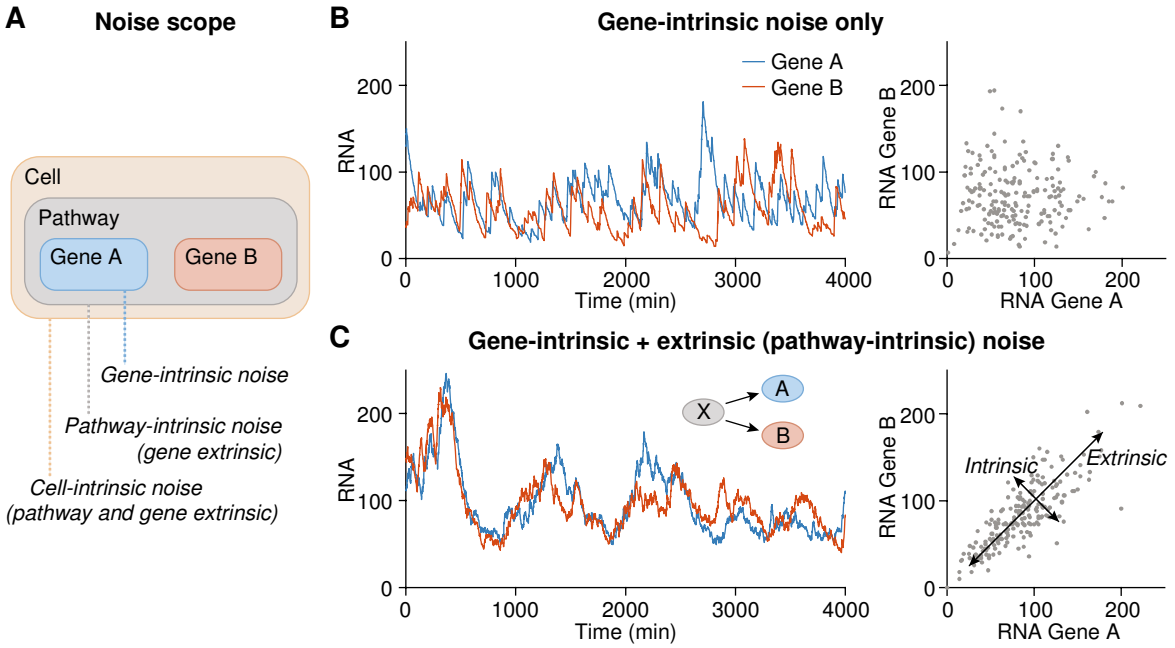

Figure 5: Intrinsic and extrinsic contributions to gene expression noise. (A) Intrinsic and extrinsic sources of noise are context dependent. Noise that is intrinsic with respect to the cell or a pathway is extrinsic with respect to single genes. **(B)** Gene intrinsic noise leads to uncorrelated bursting of two genes, even if the same kinetic parameters determine bursting. Stochastic simulations were carried out twice with the same parameters (left). Individual time points are not correlated between the two genes (right). **(C)** When extrinsic noise dominates through fluctuations in an upstream signaling molecule, correlation in the expression of downstream genes is apparent.

A major aim of this thesis is to understand how gene-extrinsic factors alter the bursting behavior of an individual gene. To this end, gene intrinsic and gene extrinsic noise was quantified from long term, live-cell imaging experiments and two alleles within the same cell were studied. A mathematical model was generated and used to analyze which burst parameters differ between cells.

1.4.4 Control and biological functions of noise

Intrinsic noise from stochastic effects is not controllable, but cells can tune burst parameters to achieve specific mean and noise levels in gene expression. When promoter transitions are much faster than mRNA half-lives, the system behaves like the “always ON” model and assumes a low-noise state. In contrast, when promoter transitions are slow, noise increases substantially (Kaern et al. 2005). By regulating transition kinetics, a cell can therefore fine-tune noise characteristics of any given gene to a desired level. When transcript levels are regulated by modulating the burst frequency, this gives rise to a characteristic noise-mean scaling. In this case, increasing transcript levels are generated through shorter inter-burst intervals, with fast promoter fluctuations efficiently buffered at constant mRNA half-live. Hence, mRNA noise decreases with mean expression and this occurs with a characteristic inverse noise-mean scaling (Singh et al. 2010, Swain et al. 2002). In contrast, expression control by burst size modulation does not alter noise levels. The mechanism of gene regulation therefore entails a specific noise scaling.

Noise on the mRNA level can also be regulated downstream of transcription, when processes that are slower than promoter switching buffer fluctuations. For example, slow nu-

clear export of mRNAs leads to more stable cytoplasmic mRNA concentrations (Stoeger et al. 2016), miRNAs can regulate protein noise (Schmiedel et al. 2015), and long mRNA and protein half-lives also reduce noise levels. Further tuning of gene expression noise occurs when genes are embedded in gene regulatory networks, in which feedback mechanisms can amplify or dampen stochastic fluctuations (Kaern et al. 2005). The multitude of mechanisms for noise control suggest that noise in gene expression can be adjusted in a gene-specific manner and is an evolvable trait itself (Fraser et al. 2004).

The non-determinism that arises from stochastic effects compromises reliable functioning and was long thought to be detrimental to cellular function, especially in unicellular organisms, where strong fluctuations would be of consequence on survival when an essential gene develops a low expression level by chance. Nevertheless, fluctuations also provide vital biological functions. In a population of unicellular organisms, noisy expression may lead to the formation of heterogeneous phenotypes with benefits in rapidly changing environments and in response to stress—a phenomenon called “bet-hedging”. For example, subpopulations are responsible for antibiotic resistance in bacteria (Balaban 2004). When cells spontaneously switch in and out of a persistent state with reduced growth but insensitivity to antibiotic treatment, this increases fitness of the whole population. In multicellular organisms, one example for the usefulness of expression noise is decision-making during differentiation. The heterogeneous expression of key transcription factors can lead to separation of lineages from otherwise identical cells. For instance, expression of *Nanog* or *Gata6* occurs in the inner cell mass of a mouse blastocyst in a mutually exclusive manner and leads to epiblast and primitive endoderm formation, respectively (Chazaud et al. 2006). Similarly, in the hematopoietic lineage, cells with extremely low or high levels of the stem cell marker *Sca-1* preferentially assume an erythroid or myeloid lineage, respectively (Chang et al. 2008). An unfavorable case of cellular heterogeneity is the escape of cancer cells from chemotherapeutic treatment (Cohen et al. 2008, Paek et al. 2016), highlighting the necessity to understand origins and control of noise in biological systems.

1.5 Experimental techniques to study single-cell transcription

Assessing population heterogeneity and noise arising from stochastic gene expression requires single-cell approaches, as it is not possible to distinguish differences in the distribution of a single-cell quantity based on population measurements. For example, when the distribution of RNA numbers in Figure 4A and B is compared, it is evident that their specific shape is determined by the underlying gene expression model. However, the population mean, identical in both conditions, does not describe the extent of distribution within the population. Several experimental techniques exist to quantify gene expression in single cells; these have differences in throughput and in the possibility to perform time-resolved measurements (dynamic vs. snapshot measurements).

Snapshot measurements assess cellular properties at a single time point for each individual cell, often involving fixation or lysis of cells. Flow cytometry is one example where cells are observed once. It allows for analysis of protein levels through the expression of fluorescent proteins or following staining with antibodies. Therefore, flow cytometric meas-

urements are ideally suited to study noise on the level of proteins (Dey et al. 2015, Newman et al. 2006). While cytometry provides a very high throughput, by characterizing millions of cells, absolute quantification is difficult. Moreover, the analysis is restricted to proteins, which only allows for indirect conclusions on promoter regulation because mRNA processing and translation act to buffer promoter fluctuations.

Another inherently single-cell method is microscopy, which has been used for decades to study cell-to-cell differences, for example, heterogeneity in the expression of a glucocorticoid-inducible β-galactosidase reporter (Ko et al. 1990). A major milestone in the visualization of single-cell transcription was the development of single molecule RNA FISH (smRNA FISH) (Femino et al. 1998). The hybridization of multiple fluorescently labeled oligonucleotides to an RNA of interest enables visualization of single transcripts as diffraction-limited spots in the cytoplasm of cells and provides the possibility to digitally count transcript abundance. Consequently, the absolute number of target transcripts can be determined, overcoming limits of measurements based on protein abundance. This technique also permits staining and quantification of nascent transcripts at the site of transcription, and hence, provides a direct measure of ongoing RNA production. smRNA FISH has been used to assess transcript heterogeneity in cell populations, with subsequent inference of kinetic parameters of transcription (Larson et al. 2013, Raj et al. 2006, Zenklusen et al. 2008). The resulting transcript distribution provides a discrimination between transcriptional bursting and constitutive transcription mechanisms (compare Figure 4A and B). Furthermore, quantification of multiple RNA species within the same cell is possible, through repetitive rounds of hybridizations (Lubeck et al. 2014) or spectral barcoding (Lubeck & Cai 2012), although multiplexing capabilities are limited. Quantification of the whole transcriptome of a single cell in an unbiased manner is enabled through recent developments in sequencing technologies (Tang et al. 2009). This allows for genome-wide noise measurements and inference of expression kinetics (Kim & Marioni 2013). However, all of these methods are snapshot measurements and share the disadvantage that a single cell can be analyzed only once. Detailed analysis of promoter kinetics, for example, resolving distributions of burst intervals (Figure 4B and C, right) requires sequential dynamic measurements in individual living cells.

Live-cell microscopy is the method of choice to observe cellular dynamics. Transcriptional bursting has been studied by time-resolved luminescence measurements based on luciferase expression (Harper et al. 2011, Suter et al. 2011), but inference of promoter kinetics from protein fluctuations is indirect and necessitates knowledge about RNA and protein stability. Direct observation of temporal fluctuations in promoter activity is possible when nascent RNAs at the transcription site are labeled with fluorescent proteins. To achieve sequence-specificity, the method exploits binding of bacteriophage coat-proteins to hairpin structures within its RNA genome (Bertrand et al. 1998, Chao et al. 2008, Fusco et al. 2003). Different systems exist based on the combination of RNA sequence and coat protein, for example from bacteriophage MS2 or *Pseudomonas* phage 7 (PP7), and their combination permits labeling of two separate RNA species (Hocine et al. 2012). The DNA that codes for hairpin-forming RNA sequences is introduced into the gene-of-interest, and

fluorescently labeled coat-proteins are co-expressed within the cell. Upon transcription, coat proteins bind to stem-loop structures in the nascent RNA (Figure 6A) and lead to an accumulation of fluorescent proteins at the site of active transcription, which becomes detectable as a fluorescent spot above the background of unbound coat-proteins (Figure 6B). As long as labeled transcripts are bound to elongating RNA polymerases, their diffusion is constrained. Transcripts diffuse away from the site of active transcription following termination, leading to lower fluorescence at the transcription site. The intensity of the spot is therefore a direct measure for the number of nascent transcripts currently being synthesized on the gene. Multiple repeats of RNA stem-loops are usually used to increase the number of recruited fluorescent proteins. This improves the signal-to-noise ratio to the point where even single transcripts can be observed as diffraction limited spots in the cell.

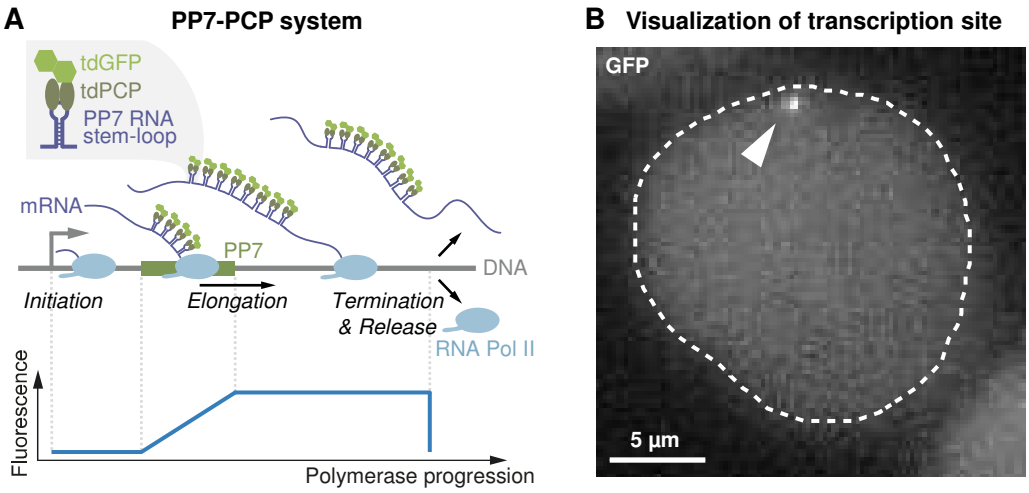

Figure 6: PP7-PCP system visualizes transcription in living cells. (A) Stem-loop structures are formed by PP7 sequences in nascent RNAs that protrude from elongating RNA Polymerases. GFP-labeled PP7 coat proteins (tdPCP-tdGFP) bind to these stem-loops and contribute to the fluorescence signal of the transcription site (below) as long as the polymerase is elongating. **(B)** Microscopic image of a transcription site (arrowhead) that is visible as a bright spot within the nucleus (dashed line). Nuclear background fluorescence is a result of nuclear localized unbound tdPCP-tdGFP.

In this thesis, the PP7 system was used to label nascent transcripts of an endogenous gene in order to study its non-genetic heterogeneity and to quantify its bursting kinetics. Furthermore, long-term observation of multiple single cells was performed to provide sufficient representation of population characteristics to discriminate intrinsic and extrinsic variability in nascent RNA production.

1.6 Nuclear hormone receptors and estrogen signaling

The modulation of gene expression by nuclear hormone receptors (NRs) is a paradigm of gene regulation in higher eukaryotes. The nuclear receptor superfamily consists of a variety of structurally related proteins (48 in humans) that act as ligand-dependent transcription factors with roles in development, cell growth, reproduction, and metabolism (Robinson-Rechavi 2003). They mediate the effect of small lipophilic signaling molecules (ligands) such as steroid hormones, retinoids, thyroid hormones, and vitamin D3, through linking these signals to a transcriptional response. Nuclear receptors bind to specific DNA

sequences (hormone response elements) in the enhancers of target gene promoters. The response elements are often found in two palindromic copies, leading to binding of homo- or heterodimers of NRs (Helsen et al. 2012). In the absence of ligand, NRs are usually bound by inhibitory proteins, leading to repression or localization of the NR to the cytoplasm. Ligand binding induces a conformational change, which leads to displacement of inhibitory proteins, recruitment of coregulatory proteins, and activation of target gene transcription. For example, HATs are recruited that acetylate histones and open the chromatin for transcription with subsequent recruitment of the basal transcription machinery. Ligand-activated repression is also possible, depending on the recruited proteins (Carr & Wong 1994).

Estrogen is the main female sex hormone and is involved in the control of sexual behavior and reproductive functions. It regulates the proliferation and differentiation of reproductive organs, principally the uterus, breast and ovaries in women, but also acts on the nervous and vascular system (Gruber et al. 2002). Estrogen is also a key determinant of bone mineral density in females. The biologically active agents are the three naturally occurring estrogens estrone (E_1), 17β-estradiol (E_2) and estriol (E_3) (Figure 7A), which are derivatives of the steroid cholesterol.

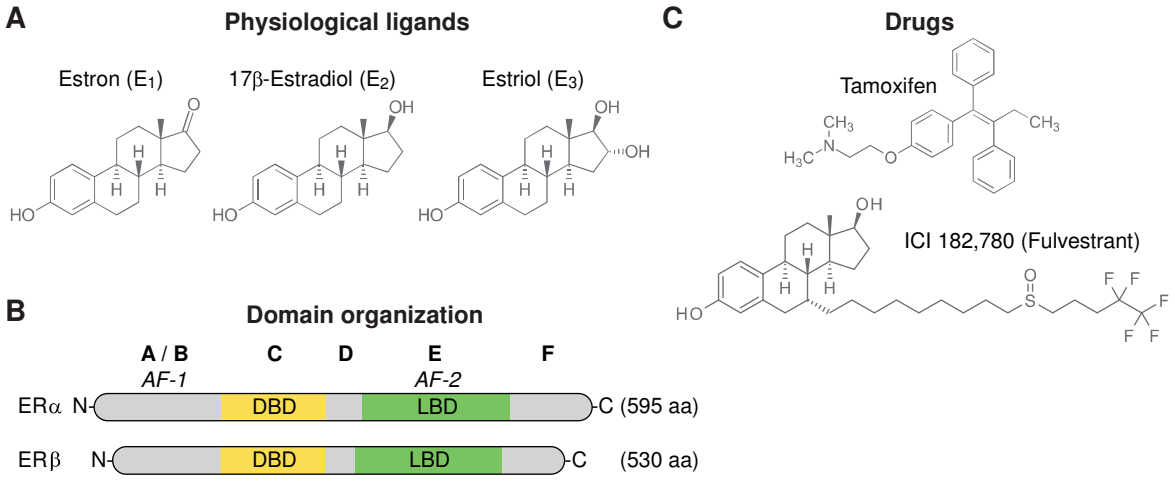

Figure 7: Structure of estrogen receptors and their ligands. (A) Chemical structure of the three naturally occurring estrogens. **(B)** Domain organization of ERα and ERβ. The N-terminus (A/B) contains the ligand-independent activation function (AF) 1. The DNA-binding domain (DBD, C) and the ligand-binding domain (LBD, E) are connected via a hinge region. **(C)** Chemical structure of two anti-estrogenic drugs that are used in cancer treatment.

Estrogens diffuse through the plasma membrane and bind to estrogen receptors (ER). The receptors exist in two main subtypes, ERα and ERβ, that share a common organization of functional domains (Figure 7B) (Lipovka & Konhilas 2016): A central DNA-binding domain (region C, 96 % amino acid identity) containing two zinc-finger motifs facilitates DNA binding and is connected via a hinge region to the ligand-binding domain (region E, 58 % identity). Liganded receptors undergo a major conformational change that displaces receptor-bound chaperones. This permits receptors to bind as dimers to estrogen response elements (EREs) in DNA and to recruit co-regulators, thereby controlling target gene expression (McDonnell & Norris 2002). In the absence of functional EREs, transcrip-

tional control is possible through “transcriptional crosstalk” with a second transcription factor (Göttlicher et al. 1998), accounting for regulation of about one third of estrogen target genes (O’Lone et al. 2004). Besides this “classical” pathway, alternative mechanisms of ER action exist, through crosstalk to other signaling pathways via receptor phosphorylation (Weigel 1996) and membrane-associated functions (Watters et al. 1997).

Dysregulation of the physiological homeostatic and proliferative roles of estrogen can lead to abnormal cell growth. Indeed, ERs have an important role in breast cancer development, with about 70 % of breast cancers hormone-dependent and ERα positive at the time of diagnosis (Lumachi et al. 2013). Two hypotheses exist how estrogen could increase cancer risk (Deroo & Korach 2006): First, an increased cell division rate driven by estrogen gives rise to a higher probability for replication errors and the risk for secondary mutations. Second, genotoxic DNA-damaging by-products of estrogen metabolism could increase mutation rates. Proliferative effects of estrogen are mediated through transcriptional regulation of target genes, among which are cell cycle regulators, including *cyclin D1*, oncogenes, such as *c-myc*, and other growth regulators, like *GREB1* (Yamaga et al. 2013). Several endocrine therapeutic strategies are available to treat ERα positive breast cancer, which either target estrogen production through ovarian suppression or aromatase inhibitors, or target ERα directly (Lumachi et al. 2013). Selective estrogen receptor modulators (SERMs) such as tamoxifen, or ER downregulators like fulvestrant (ICI 182,780) (Figure 7C) are used as receptor agonists, or to reduce receptor protein levels, respectively.

Treatment strategies improved the prognosis of ERα positive breast cancers as compared to tumors lacking ERα expression. Considering the role of ERα-mediated gene expression in the development and prognosis of breast cancer, it is important to understand the molecular mechanism of transcriptional regulation and the effects of therapeutic intervention. Given the spatial and temporal heterogeneity within carcinomas (Martelotto et al. 2014), it will become increasingly important to understanding genetic and non-genetic origins of heterogeneity. The aims of this thesis were to monitor estrogen-regulated transcription, its associated cell-to-cell variability, and to understand physiological regulation and the effect of drug treatment on the expression of a gene involved in growth regulation.

1.7 Chromatin dynamics in the estrogen response

Estrogen-dependent activation of target genes involves remarkable dynamics on the level of chromatin modification and protein association at the promoter. The promoter of the estrogen-dependent *pS2* (trefoil factor 1, *TFF1*) gene has been studied extensively and has unraveled ordered and sequential processes that are necessary for productive transcription to occur. In a landmark study, Métivier and colleagues (2003), based on a previous study of Shang et al. (2000), characterized the dynamics and co-occupancy of 46 factors upon estrogen stimulation. This included ERα, enzymes that modify chromatin, transcription factors, the transcription machinery plus histone modifications and chromatin remodeling. Remarkably, an orchestrated recruitment and displacement of proteins to the

promoter occurs in a cyclical fashion, with duration of 40–60 minutes (Figure 8) along with covalent chromatin modifications.

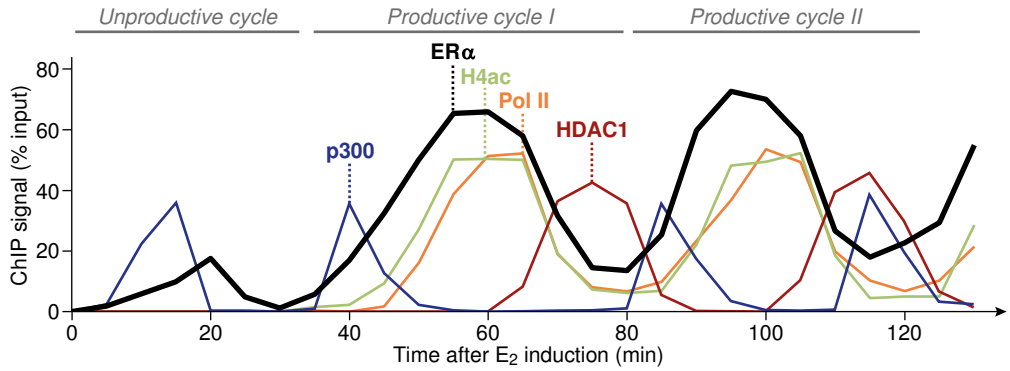

Figure 8: Chromatin dynamics at the *pS2* promoter upon estrogen stimulation. Cyclical promoter binding of ERα is observed in a synchronized population of cells upon release of estrogen starvation. An initial unproductive cycle is due to the synchronization process and prepares the promoter for activation. The HAT p300 acetylates histones (H4ac) at the beginning of each productive cycle, and polymerases are recruited for transcription. Histone deacetylation through HDACs occurs at the end of each cycle. The data for this figure was taken from (Métivier et al. 2003).

In a series of ordered events, ERα recruits coactivators that lead to the assembly of the transcription machinery and initiation of transcription. At the beginning of a transcription cycle, chromatin accessibility is increased through recruitment of the SWI/SNF (Switch/Sucrose Non-Fermentable) chromatin-remodeling complex. HATs and HMTs are recruited to acetylate and methylate histone residues, and DNA is demethylated, which collectively promotes a transcriptionally permissive promoter state. Subsequent recruitment of general transcription factors and the transcription machinery leads to polymerase binding and activation. At the end of a productive cycle, the promoter is cleared from transcription factors in a proteasome-dependent manner (Reid et al. 2003), chromatin is deacetylated and demethylated, DNA is remethylated, and the SWI/SNF complex restores the original nucleosome organization. While the *pS2* promoter is by far the best studied example, cyclical TF association and chromatin modifications were also observed for other ERα target genes, for example *Wisp-2* (Métivier et al. 2008), *cyclin D1* (Park et al. 2005), or *c-myc* (Shang et al. 2000). It is thought that such periodic limitation of transcription allows for faster response to changes in stimulus conditions, as the cell has to continuously reactivate transcription according to the presence of hormone (Carlberg 2010, Rybakova et al. 2015).

It is tempting to speculate that such chromatin-mediated periodic transcriptional permissiveness as observed in cell populations would provide a mechanism for the occurrence of transcriptional bursts in single cells. Interestingly, a simple two- or three-state promoter model cannot describe the regular oscillations with almost no dampening. Instead, a large number (>100) of promoter states is necessary (Lemaire et al. 2006), in agreement with the multitude of biochemical events at the promoter chromatin. By observing transcriptional bursting of an estrogen-dependent promoter, this thesis tries to understand the relationship between population-level chromatin dynamics and single-cell stochastic transcription.

1.8 Aims

Gene expression and signal responses vary between cells due to stochastic effects. While variegation can be beneficial in tissue with heterogeneous cell populations, it is also exploited by cancerous tissue, with intra-tumor diversity lowering therapeutic effectiveness and consequently promoting drug resistance. To gain a deeper understanding into the origins and consequences of cellular heterogeneity, this thesis aims to characterize the expression of a key growth regulator, *GREB1*, that is regulated by estrogen. Three main questions are addressed:

1) How does estrogen regulate transcriptional output of target genes in the context of transcriptional bursts? There is evidence that enhancers regulate the frequency of transcriptional bursts (Bartman et al. 2016, Fukaya et al. 2016) and similarly, this has been observed for a nuclear receptor-controlled gene (Larson et al. 2013). However, evidence is lacking for an endogenous locus under the control of physiological signaling.

2) How does cellular state, i.e. extrinsic noise, fine-tune the transcriptional response? Differences in the concentrations of regulatory proteins and metabolites as well as the cellular size and microenvironment impinge on gene expression (Battich et al. 2015), but how these factors affect burst kinetics and multiple alleles remains unclear.

3) What is the influence of chromatin on bursting and transcriptional heterogeneity? Chromatin seems to influence burst characteristics (Lenstra et al. 2016), but the effects are likely gene-specific. By studying the estrogen response, in which chromatin dynamics are particularly well described, a link to noise in single cells might be established. Small-molecule inhibitors of epigenetic regulators will be used as additional perturbation.

In order to answer these questions, I aim to image nascent transcripts in living cells under various conditions. A knock-in strategy will be used to label an endogenous gene within an unperturbed chromatin environment. Acquired datasets will inform a kinetic model of promoter progression. Model fitting will then be employed to discriminate between alternative hypotheses and to motivate experiments that challenge their prediction. Ultimately, I aim to establish a unifying model of estrogen-dependent transcription, which quantitatively incorporates estrogen-dependency and cell-to-cell variability. This work will provide novel insight into the role of non-genetic variability of gene expression in heterogeneous cancer growth.

2 Results

2.1 Generation of cell lines to visualize endogenous estrogen-dependent transcription

2.1.1 Design criteria for reporter cell lines

This thesis centers on the quantification of estrogen-dependent transcription through direct visualization of transcriptional activity in living cells at the level of a single allele. The PP7 system was utilized to visualize nascent transcripts by fluorescence microscopy. This approach relies on the incorporation of stem-loop forming sequences into the genome and co-transcriptional binding of GFP-labeled PP7 coat proteins (GFP-PCP) to nascent RNAs. PP7 sequences were used because they achieve almost complete occupancy of stem loops with coat-proteins, as compared with a system that is based on sequences from the bacteriophage MS2 (Wu et al. 2012). A reporter cell line was created by knocking-in PP7 sequences into the genome of MCF-7 (Michigan Cancer Foundation-7) human breast cancer cells (Soule et al. 1973). This cell line expresses ERα and is a well-established model system for estrogen-induced transcription (Lee et al. 2015). The reporter cell line was supposed to enable visualization of endogenous estrogen-controlled transcription. Therefore, a well-characterized, estrogen-dependent gene was chosen for PP7 labeling. An ideal target gene has a high level of expression, such that initiation of transcription is observed at a high frequency. Furthermore, a long transcriptional unit is desirable, because it provides a large dwell-time during transcription, i.e. the time that the transcript is associated with the gene while the polymerase is elongating. This would increase the observation time for each nascent RNA and consequently, the intensity of transcription foci.

I chose *growth regulation by estrogen in breast cancer 1* (*GREB1*) as a target gene for visualization of transcription. GREB1 is an important mediator of estrogen-induced growth in breast cancer cells (Rae et al. 2005) and acts as an estrogen-specific cofactor of ERα (Mohammed et al. 2013). Consequently, it represents a biologically relevant target gene to study transcriptional heterogeneity and the effects of estrogen treatment. Furthermore, it fulfills the above-mentioned selection criteria for visualization: the promoter/enhancer region contains at least three estrogen response elements (EREs) that directly bind ERα (Deschênes et al. 2007, Sun et al. 2007). *GREB1* is among the highest expressed estrogen-induced genes (Ghosh et al. 2000). Its transcriptional unit has 33 exons that span 109 kb of genomic DNA, resulting in an elongation time for the whole gene of ~30 minutes when an elongation rate of 3–4 kb/min is assumed. Such kinetics seem plausible according to data from kinetic polymerase occupancy measurements along the gene (wa Maina et al. 2014 and personal communication with Magnus Rattray).

Introducing PP7 sequences into a gene can have adverse effects on its function. Because *GREB1* is important for cell growth, it was important to affect its activity only minimally. Within the coding region, PP7 sequences would introduce premature stop codons and disrupt protein structure. Therefore, I introduced PP7 sequences within the untranslated regions (UTRs) of *GREB1* (Figure 9A, top). Both, the 5' UTR in exon 2, as well as the

3' UTR in exon 33 were chosen. Because changes within the UTR sequences may alter RNA stability, I also chose to introduce PP7 sequences into intron 2, such that they would be removed from the pre-mRNA and leave an unaltered mRNA.

The transgene was designed to introduce an array of 24 PP7 sequences into the nascent *GREB1* transcript, along with a selection cassette to aid isolation of clones with a successful knock-in (Figure 9A, middle). The selection cassette consists of a strong viral promoter driving expression of a puromycin resistance gene and of blue fluorescent protein. This allowed selecting cells that integrated the transgene by puromycin treatment or by fluorescence. I designed the selection cassette such that it is expressed from the antisense strand, as otherwise the polyadenylation signal would terminate *GREB1* transcription prematurely. Furthermore, the selection cassette was flanked with loxP sequences that enable Cre recombinase-mediated excision, to produce a *GREB1* locus that enables visualization of transcription but does not contain more artificial sequences than necessary (Figure 9A, bottom). Such a minimally altered *GREB1* locus is likely to maintain endogenous regulation and transcriptional behavior.

2.1.2 Knock-in of PP7 sequences into the *GREB1* gene

Knock-in cell lines were generated as outlined in Figure 9B. The knock-in was performed by transient expression of the nuclease Cas9 with a specific guide RNA in the presence of a co-transfected DNA template for homology-directed repair of the double strand break (Cong et al. 2013, Yang et al. 2013). Guide RNAs and repair templates were matched based on the desired knock-in location. Clonal cell lines were derived using puromycin selection and genotyped by PCR on genomic DNA (Figure 9C). Later analysis (see 2.1.3) revealed that clonal cell lines carry a knock-in within a single allele of *GREB1* for exon 2 and exon 33, while the knock-in for intron 2 occurred in two alleles. Stable co-expression of the GFP-fused PP7 coat protein was achieved through random genomic integration by Sleeping-Beauty transposase (Mátés et al. 2009) and isolation of a clone with low GFP fluorescence.

A single clone harboring the knock-in within exon 2 was selected for transient expression of Cre recombinase, to excise the selection marker. Single cells with successful excision were isolated by flow cytometry, based on the absence of BFP fluorescence, and the recombination event was confirmed by genomic PCR (Figure 9C). The resulting cell line, MCF7-PCP_GREB1_ex2_c16_Cre, was used for most of the live-cell experiments throughout this thesis.

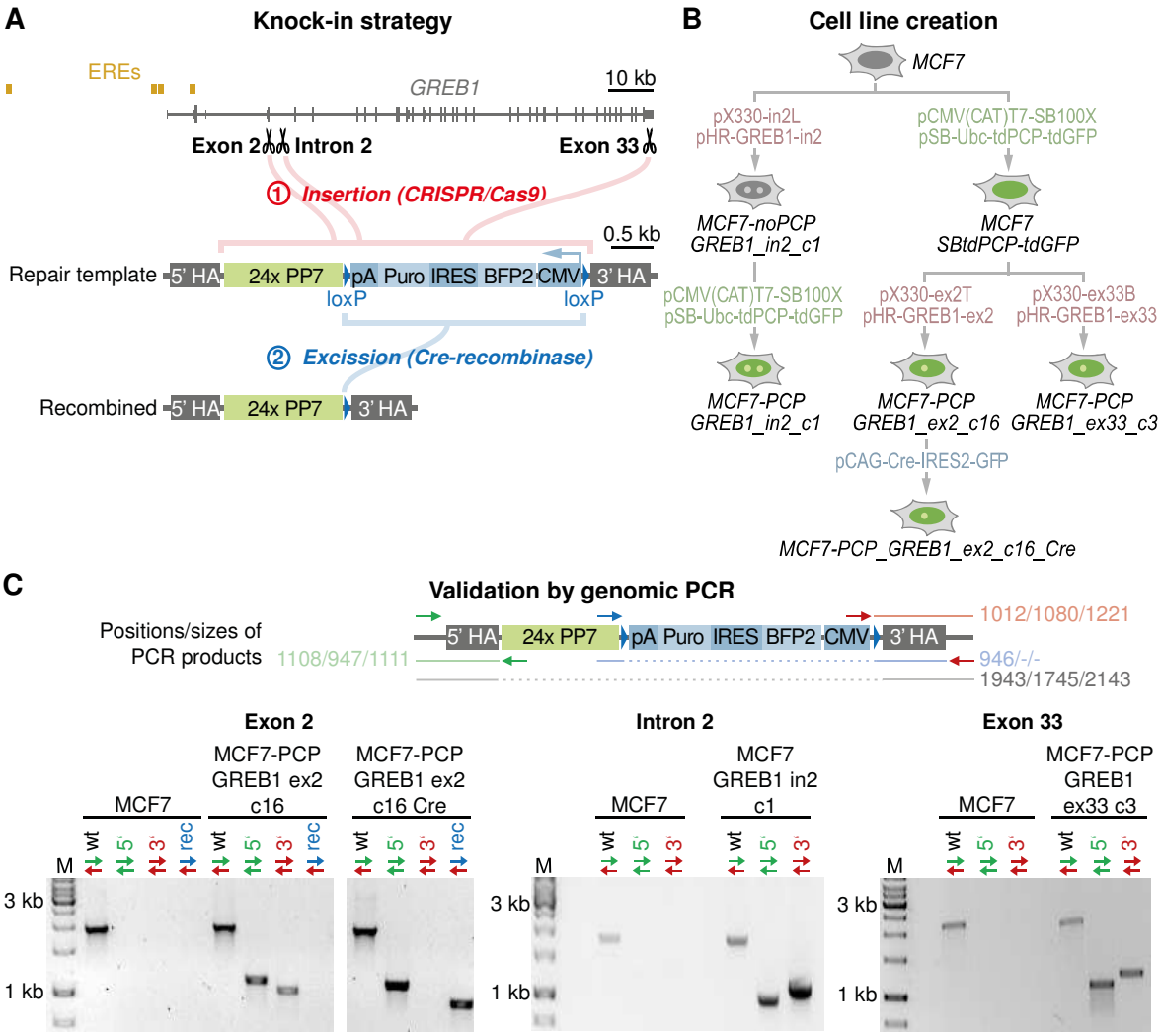

Figure 9: Creation of knock-in cell lines and validation of genome engineering. (A) Strategy to knock-in PP7 sequences into *GREB1*. An array of 24x PP7 sequences was introduced together with a floxed selection cassette into three different locations within *GREB1* (scissors). The selection cassette was excised with Cre-recombinase. (ERE: estrogen response element, HA: homology arm, pA: polyadenylation site, Puro: Puromycin resistance, IRES: internal ribosomal entry site, CMV: promoter of cytomegalovirus). **(B)** Workflow of cell line creation. The order of steps differed between cell lines as outlined. Each arrow stands for transfection with the indicated plasmids (green: tdPCP-tdGFP, red: CRISPR/Cas9, blue: Cre-recombinase) and creation of a clonal cell line (for details see methods). A green nucleus represents expression of tdPCP-tdGFP. A bright spot in the nucleus indicates knock-in of PP7 sequences. Note the two integration sites in the GREB1_in2_c1 clone. **(C)** Validation of genome engineering by PCR on genomic DNA. The location of primers and the size of PCR products (in bp) is indicated in the scheme above (order: exon 2/intron 2/ exon 33). Primers outside the homology arms are specific to the knock-in location, hence, their product size depends on the cell line. (wt: wildtype locus, 5': correct 5' recombination, 3': correct 3' recombination, rec: successful Cre-mediated recombination, M: 1kb DNA ladder).

2.1.3 Visualization of *GREB1* transcription sites

When GFP-PCP was co-expressed in the knock-in cell lines, nascent *GREB1* transcripts became apparent as a bright fluorescent spot within the cell nucleus in living cells (Figure 6B). I wanted to assess whether the observed spots represent actual sites of ongoing transcription and can be used to quantify *GREB1* transcription. To increase statistical power of my analysis, I quantified the occurrence and fluorescence intensities of spots in thousands of cells by high-content microscopy. This required fixation of cells prior to imag-

ing, allowing snapshot analysis at a single time point. The GFP signal is retained after fixation (Figure 10A) and spots could be automatically identified and quantified from microscopic images (Figure 10B).

Cells were grown at 100 pM E_2 and treated with either of two transcriptional inhibitors, Actinomycin D and Flavopiridol. Both drugs reduced spot intensities such that almost no spot was detectable. Hence, the observation of spots is dependent upon transcription. In addition, treatment with the anti-estrogens ICI 182,780 and 4-Hydroxy-Tamoxifen (OHT) showed reduced spot intensities, confirming that estrogen signaling is necessary for transcription sites to appear. The results of the inhibitor treatments confirmed that nuclear GFP spots result from estrogen-dependent transcription and that the automatic image analysis reliably quantified spots in fixed cells. Below, I used the same image analysis pipeline to study the effect of different E_2 concentrations on transcription of *GREB1* (see 2.1.4).

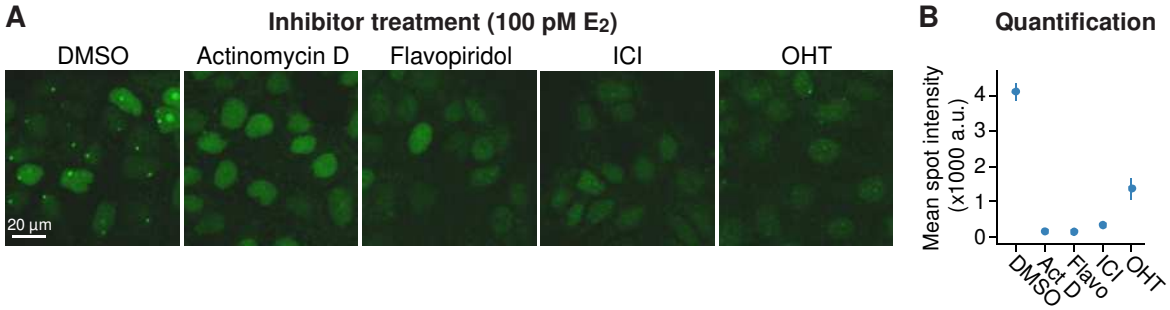

Figure 10: Occurrence of nuclear spots is dependent upon transcription and estrogen signaling. **(A)** Images of MCF7-PCP_GREB1_ex2_c16_Cre cells grown at 100 pM E_2 and treated with DMSO (solvent control), with inhibitors of transcription (Actinomycin D or Flavopiridol) or inhibitors of estrogen signaling (ICI 182,780 or 4-Hydroxy-Tamoxifen (OHT)). Spot intensities are reduced upon inhibition of estrogen signaling or transcription. **(B)** Intensity quantification of automatically detected spots. Mean and standard deviation over all cells of two independent replicates is shown.

To confirm that the accumulation of GFP fluorescence indeed occurs at the endogenous *GREB1* locus, I performed single-molecule RNA fluorescence *in-situ* hybridization (smRNA FISH) as an independent experimental method (Figure 11). This method relies on the hybridization of multiple fluorescently-labeled oligonucleotides to the RNA of interest and allows visualization of transcripts with single-molecule resolution (Raj et al. 2008). Probes for exonic and intronic regions of *GREB1* with distinct fluorescent dyes were hybridized simultaneously.

Exonic probes labeled mature transcripts in the cytoplasm, which became visible as diffraction-limited spots in the microscope. Transcription sites were apparent as bright foci in the nucleus, as many nascent transcripts co-localized at this position. These transcription sites were also visible when probes for intronic regions were used, as unspliced introns are also present at this locus. The FISH signal for intronic and exonic *GREB1* probes co-localized at high-intensity transcription sites confirming sequence specificity of the hybridized probes.

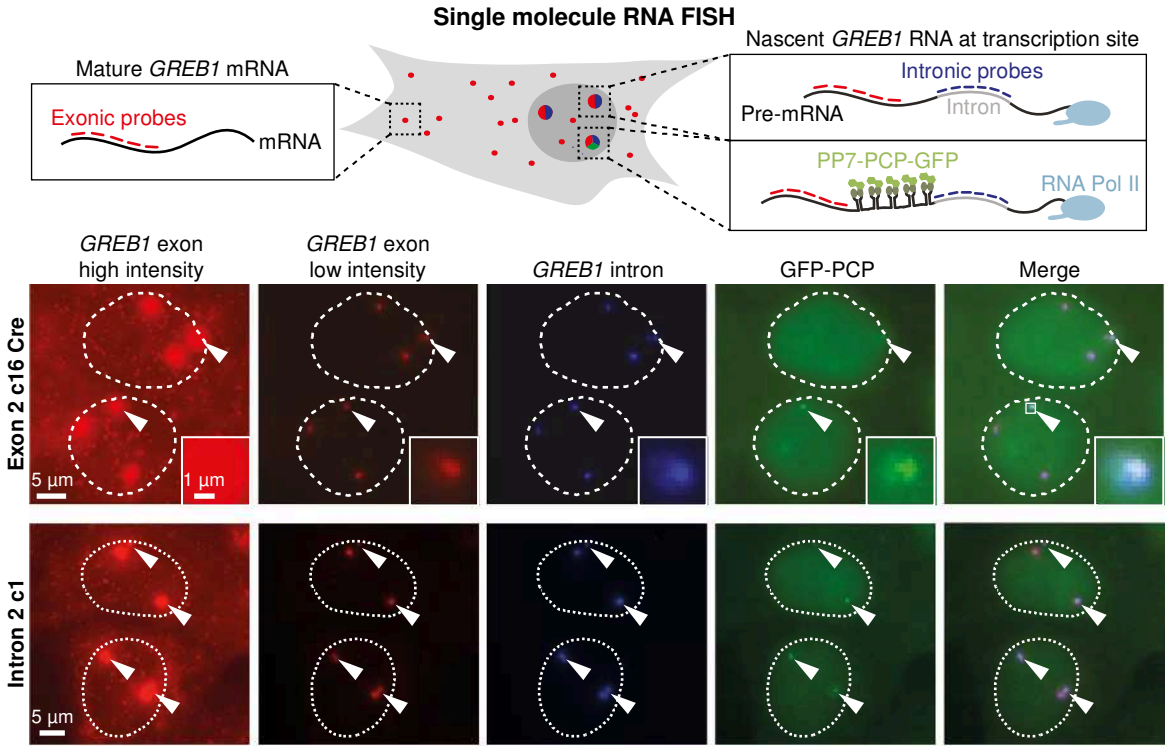

Figure 11: PP7-PCP signal co-localizes with *GREB1* smRNA FISH foci. MCF7-PCP_GREB1_ex2_c16_Cre and MCF7-PCP_GREB1_in2_c1 cells were grown at 100 pM E_2, fixed, and hybridized with exonic and intronic probes for *GREB1*. Exonic probes label mature transcripts that are visible as diffraction-limited spots. Transcription sites are evident as bright exonic foci in the nucleus (dashed lines) that co-localize with foci of intronic probes. One of the three transcription sites shows GFP accumulation (arrowheads) due to the presence of knocked-in PP7 sequences in the exon 2 c16 Cre cell line. The PP7-labeled transcription site of the lower cell is shown in the inset. Co-localization is apparent at two out of three loci in intron 2 c1 cells.

Up to three bright foci in the nucleus were observed, suggesting that the *GREB1* gene is present at three copies in the genome of MCF-7 cells. As anticipated, MCF7-PCP_GREB1_ex2_c16_Cre cells had one transcription site that also co-localized with the spot visible in the GFP channel originating from the PP7-PCP labeled locus. Thus, single-molecule RNA FISH revealed that one out of three endogenous *GREB1* loci was successfully labeled with PP7 sequences and that the spots in the GFP channel coincide with the *GREB1* locus. Hence, this cell line can be used to quantify nascent *GREB1* transcription. In MCF7-PCP_GREB1_in2_c1 cells, up to two transcription sites were visible as bright spots in the GFP channel, with both of them co-localizing with exonic and intronic foci in smRNA FISH images.

2.1.4 Location of PP7 stem-loops characterizes transcriptional kinetics

I wanted to assess whether the location and properties of the knock-in sequences influence characteristics of the observed transcription foci. The fluorescence intensity of the spot is a direct measure for the density of polymerases that actively transcribe the gene. Because transcripts are only detected after transcription of the PP7 region, the intensity of transcription foci is dependent on the position of the stem-loop cassette within the gene (Figure 12A). Location within exon 2 leads to a long duration (~ 30 min) of observability for each RNA and hence, multiple polymerases will be observed at the gene, leading to bright spots. Labeling in exon 33 produces shorter RNA dwell-times with lower spot intensities,

and intron 2 labeling would also lead to reduced dwell-times when the intron is spliced co-transcriptionally.

I wanted to experimentally test these predictions and further assess the E_2-dependency of *GREB1* transcription. Thus, I quantified the occurrence and fluorescence intensities of transcription sites in fixed cells by high-content microscopy at various E_2 concentrations, ranging from E_2 starvation to full induction.

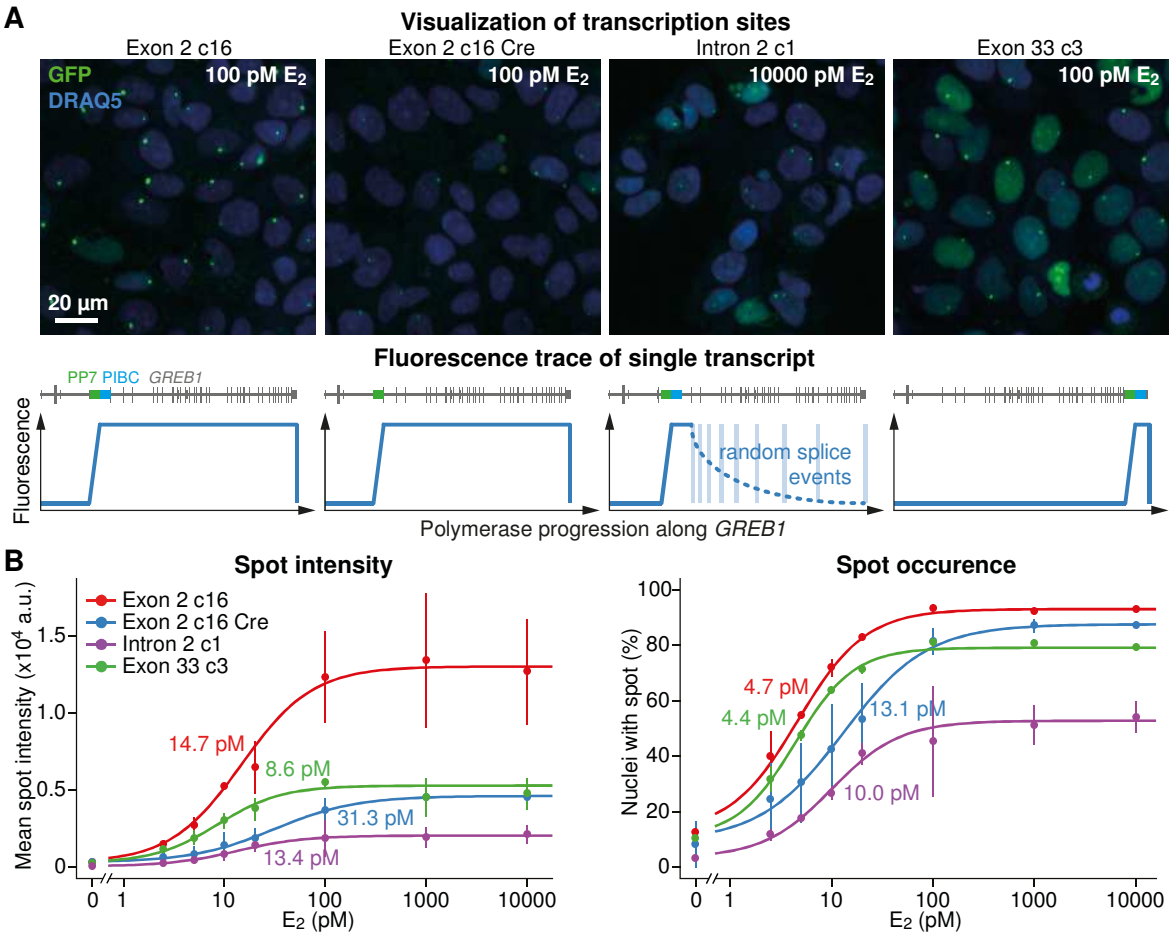

Figure 12: Observation and quantification of transcription sites in knock-in cell lines. (A) Transcription sites are visible as bright spots in the GFP channel within DRAQ5 stained nuclei. Different knock-in cell lines were grown for three days at the indicated E_2 concentrations, fixed, and imaged. A cropped region from maximum intensity projections of z-stacks are shown. A single spot is visible per nucleus for the cell lines Exon 2 c16, Exon 2 c16 Cre and Exon 33 C3, while the intron 2 c1 cell line, which has two labeled alleles, shows up to two spots. Schematic fluorescence traces of a single transcript are shown below for each of the four cell lines. Exon 2 is transcribed early, leading to long dwell-times. Intron 2 is potentially spliced co-transcriptionally with random timing, leading to decreasing intensities along the gene (dotted line = average of many transcripts). Exon 33 is transcribed late, leading to short dwell-times. **(C)** Dose-dependency of transcription sites. Spots were automatically detected and quantified from microscopic images. The average intensity of the brightest spot per cell (left) and the percentage of cells with a visible spot (right) is shown. Mean intensity and occurrence of the two most intense spots is shown for the intron 2 cell line. The EC_{50} of a fitted four-parameter Hill equation is indicated. Error bars denote standard deviation over all cells from two independent experiments.

Transcription sites were visible in all four cell lines at high concentrations of E_2 (Figure 12A) and image analysis showed that spot intensities and the fraction of cells with an observable spot increase with E_2 in a dose-dependent manner (Figure 12B). Without estro-

gen, about 10 % of cells showed a visible spot, with this value increasing to up to 90 % at saturating E_2 concentrations, highlighting an appropriate dynamic range of the experimental system. All knock-in locations showed a comparable sensitivity for E_2 with an EC_{50} of about 10 pM, while the excision of the selection cassette from exon 2 led to a slightly larger value.

The mean intensity of transcription sites recapitulated predictions based on the proposed single-transcript traces in Figure 12A. Knock-in within exon 2 led to highest spot intensities, about three-fold higher than for exon 33 and about six-fold higher than when intron 2 is labelled. The low intensities that were observed when the PP7 sequences are located within intron 2 indicate that their dwell-time is much shorter than for the proximate exon 2. This in turn suggests that intron 2 is primarily spliced co-transcriptionally. The knock-in site within exon 33 is about 17-times closer to the end of the gene than exon 2. This ratio should also be reflected in the spot intensities when the dwell-time is purely influenced by polymerase elongation. The fact that the intensities for exon 33 were only three-fold lower compared to exon 2 suggest that either the very 3' end of the gene is transcribed slowly, or that the transcripts are immobilized at the transcription site for processing after elongation. These results highlight the power of nascent RNA labeling to uncover kinetics of co-transcriptional processes, even from snapshot measurements in fixed cells.

I suspected that the presence of the selection cassette, especially the CMV promoter, within *GREB1* might affect transcriptional dynamics. Transcription sites from exon 2 labeled cells with and without selection cassette showed different intensities: the absence of a selection cassette led to about three-fold lower fluorescence intensities (Figure 12B left, red and blue lines). Whether this is due to changes in expression, that is, burst size or burst frequency, or whether elongation kinetics are altered cannot be distinguished from imaging experiments in fixed cells. Analysis of mRNA levels (see 2.1.5), however, suggests that expression levels are comparable for both knock-in constructs. The observed difference in transcription site intensities may therefore reflect longer transcript dwell-times when the selection cassette is present. I speculate that polymerases might interact with opposing polymerases that transcribe the selection cassette from the anti-sense strand, leading to reduced elongation kinetics. In addition to decreased spot intensities, the absence of the selection cassette leads to an increase in EC_{50} for E_2, emphasizing the influence of the cassette on transcription. The *GREB1* locus without selection cassette contains less artificial DNA sequence and behaves more similar to the wildtype locus, which I further confirmed below by RT-qPCR and smRNA FISH measurements.

2.1.5 Estrogen sensitivity is unperturbed in knock-in allele

I evaluated whether knock-in into exon 2 would alter the expression and inducibility of *GREB1* at the steady-state RNA level and wanted to assess the influence of the selection cassette. Therefore, I measured *GREB1* mRNA by RT-qPCR at different E_2 concentrations, with primers that are specific for either the wildtype allele or the knock-in allele, and compared their levels within the same sample. E_2-sensitivity was unaltered for the knock-in allele, irrespective of whether the selection cassette was excised or not: the EC_{50} for E_2

was about 8 pM for all *GREB1* loci (Figure 13A) and this value is comparable to the EC_{50} inferred from microscopic analysis (Figure 12B). The maximum expression level from the modified *GREB1* alleles was about three-fold lower than from the wildtype locus (4 % and 12 % of *GAPDH*, respectively), independent of the presence of the selection cassette (Figure 13A). This difference can be explained in part by the presence of three *GREB1* gene copies in the MCF-7 genome, two of which are unmodified. Furthermore, reduced transcript stability due to the presence of PP7 stem-loop structures in the 5' UTR may lead to lower steady-state mRNA levels. Sensitivity for E_2 and mRNA levels were not influenced by the presence of the selection cassette. However, there was a difference between the two cell lines at 0 pM E_2 (Figure 13A, insets) with the construct containing the selection cassette resulting in about 3-fold higher basal expression than the construct without the cassette. A possible explanation is that the strong viral CMV promoter within the selection cassette leads to an open chromatin environment in proximity to the *GREB1* promoter, facilitating transcription even in the absence of E_2.

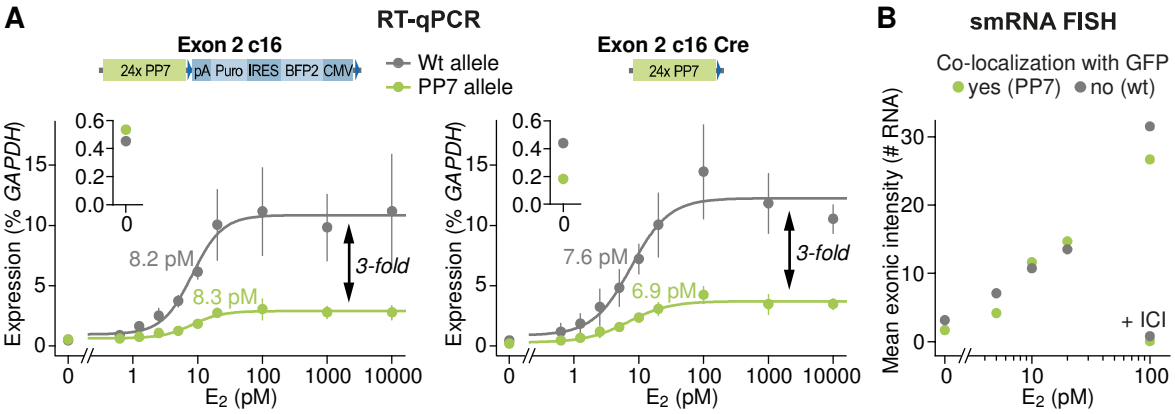

Figure 13: Knock-in allele maintains E_2-sensitivity and RNA production of wildtype locus. (A) Allele-specific RT-qPCR measurements for *GREB1* wildtype (wt, grey) and knock-in (PP7, green) sequences were performed after 18 hours of E_2 induction. EC_{50} from a fitted Hill-equation is indicated and revealed unaltered E_2-sensitivity of the knock-in allele. The inset shows expression differences at 0 pM E_2, highlighting higher leaky expression when the selection cassette is present (left). **(B)** E_2 dose-dependence of nascent transcription. Nuclear smRNA FISH signals for *GREB1* exons were quantified in MCF7-PCP_GREB1_ex2_c16_Cre cells across five E_2 concentrations. The mean intensity of the two brightest nuclear foci not co-localizing with a GFP spot (wildtype alleles, grey) is comparable to the mean intensity of the brightest focus co-localizing with GFP (knock-in allele, green) across E_2-concentrations and upon addition of 1 µM ICI 182,780 (+ ICI). This indicates that nascent transcription is unaltered by the presence of PP7 sequences.

The difference in transcript levels between wildtype and knock-in locus prompted me to analyze whether this is due to altered nascent transcription. Therefore, I quantified smRNA FISH spot intensities from exonic *GREB1* probes (Figure 13B). Bright nuclear foci represent sites of nascent transcription and allowed differentiating between wildtype and knock-in loci based on the co-localization with GFP spots from the PP7 system (see Figure 11). If PP7 sequences do not alter the *de novo* production of RNAs, the intensities from both alleles should be comparable. Indeed, I observed no differences in spot intensities across five different concentrations of E_2, confirming that nascent transcription is not disturbed by the presence of the PP7 cassette. The differences in steady-state RNA levels from RT-qPCR measurements are therefore likely due to differences in RNA stability.

Taken together, RT-qPCR and smRNA FISH experiments demonstrated that nascent transcription and E_2-inducibility of *GREB1* are not affected by the presence of PP7 sequences. The presence of the selection cassette, however, caused higher leaky expression in the absence of E_2 and confirmed detrimental effects of the additional DNA sequence containing a strong viral promoter. Potential changes in RNA stability are not crucial for measurements of transcriptional dynamics, the main observable in this study. The engineered cell line MCF7-PCP_GREB1_ex2_c16_Cre is therefore an excellent tool to study estrogen-dependent transcription in single living cells.

2.2 *GREB1* is preferentially transcribed at the nuclear periphery

The organization of the genome within the nucleus is non-random and different mechanisms are proposed that link transcriptional activity with nuclear positioning (Parada et al. 2004). However, there is no agreement on whether radial positioning of the gene influences its expression. *GREB1* transcription sites seemed to be located in close proximity to the nuclear periphery (see Figure 11 and 12A). I therefore wished to quantify this effect and additionally determine, whether the positions of spots change with induction level.

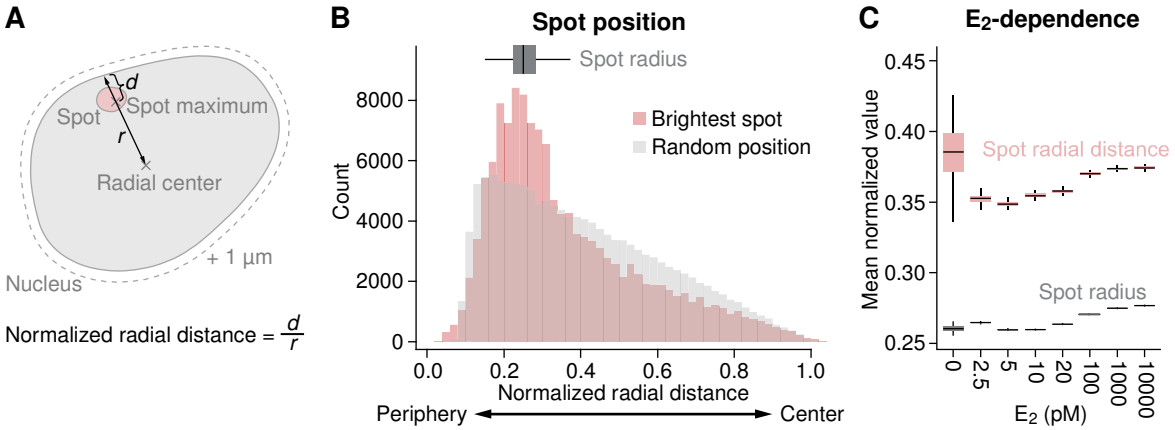

Figure 14: Transcription sites are preferentially located at the nuclear periphery. (A) Radial spot position was determined from the distance between the brightest pixel and the periphery. Spots were detected in a nuclear region that was increased by one micrometer. **(B)** Distribution of transcription site positions (red) and random positions in a circle with the same radius (grey). The boxplot above shows the normalized spot radius, indicating the closest possible peripheral localization. Preferential peripheral positioning of transcription sites is apparent. **(C)** Mean normalized spot position and spot radius as a function of E_2. Boxplots denote distributions from bootstrapping (box = 25 % to 75 % percentile, line = median, whiskers = 1.5x interquartile range).

To this end, I used the images of the high-content imaging experiments and calculated the distance from the pixel with the maximum intensity within the transcription site to the closest edge of the segmented nucleus (Figure 14A). The distribution of distances was different from that of randomly chosen positions within the nucleus, with the majority of spots being located closer to the nuclear periphery than would be expected at random (Figure 14B). Furthermore, most spots reside in a distance that is close to the spot radius, therefore, representing a location that is as close to the nuclear periphery as possible.

I wondered whether the radial position of spots changes with transcriptional activity and thus, divided the dataset according to the E_2 concentration (Figure 14C). The mean nor-

malized radial spot distance slightly increased with estrogen, suggesting that transcriptional activity could correlate with a spot position further away from the nuclear periphery. However, at the same time, the mean normalized spot radius increased because spot intensities also increase with E_2 and brighter spots usually have a bigger radius. This trend explained the apparent change in spot localization. Hence, transcription does not alter peripheral positioning. Taken together, transcription sites are preferentially located close to the nuclear periphery, and their positioning is independent of their transcriptional activity. Yet, I cannot rule out differential localization of transcriptionally inactive loci, as these are invisible to the method.

2.3 Digital modulation of transcription by estrogen

The distribution of RNA counts at the transcription site characterizes GREB1 expression heterogeneity and allows for conclusions on gene regulation. Therefore, I inferred the distribution of nascent RNA intensities at different concentrations of E_2 from high-content imaging and smRNA FISH experiments (Figure 15). In both cases, a bimodal distribution is apparent, in which transcription sites either are absent or have high intensities that result from multiple nascent transcripts. Intermediate transcript occupancies are rarely observed. The proportion between ON- and OFF-regimes changed with E_2 such that without estrogen only about 15 % of cells had a visible spot, while at saturating induction, a spot was observed in about 85 % of cells. These values are very similar between high-content imaging and smRNA FISH experiments. Furthermore, the intensities of active transcription sites increased with E_2. This analysis revealed a strong digital modulation of transcription, in which the proportion of cells showing active transcription increases. Furthermore, a weak analog (gradual) increase in transcription sites intensity for all cells is apparent.

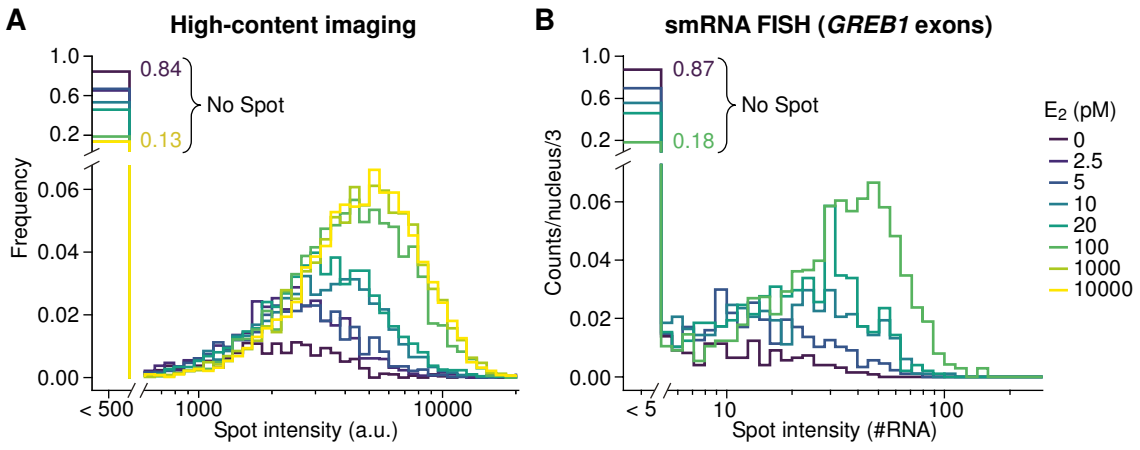

Figure 15: Distribution of spot intensities demonstrates digital modulation of transcription. **(A)** Histograms of spot intensities from high-content imaging were computed for all eight E_2 concentrations. The fraction of cells with spots, as well as the spot intensity increases with E_2. **(B)** Histograms of RNA numbers of the three brightest foci per nucleus from smRNA FISH with probes against *GREB1* exons. Counts were normalized to the number of observed nuclei in each condition. Both experimental conditions deliver comparable intensity distributions for nascent RNAs.

The observation of bimodality in the distribution of transcription site intensities would not have been possible in population measurements, which only reveal the mean of all cells. The distribution of nascent transcripts suggests that the timing of transcriptional activity is altered by E_2, for example by modulating the duration of transcriptionally silent episodes in between transcriptional bursts. However, this analysis cannot distinguish whether some cells always transcribe while others are never producing RNAs, or whether a dynamic equilibrium exists, as the transition between active and inactive phases is not directly observed. Continuous live-cell imaging can resolve this question by enabling observation of transcriptional dynamics at multiple time points within the same cell.

2.4 *GREB1* is transcribed in stochastic bursts

Snapshot measurements within single cells can be used to infer kinetic information about transcription, for example whether a gene is transcribed continuously or in bursts (Raj et al. 2006). However, time-resolved measurements contain more information and can discriminate models of gene expression on a much finer level. I wished to distinguish models of promoter progression with different complexity and analyze how individual cells differ in estrogen-dependent transcription due to differences in their internal cellular state. Time-resolved measurements are essential to perform such analyses and careful calibration of measurements is needed to allow for absolute quantification.

2.4.1 Calibration of spot intensities for absolute quantification

I wanted to quantify nascent transcription in absolute terms, i.e. derive the number of RNAs that are currently being transcribed at the *GREB1* locus from the measured fluorescence intensity of the transcription site. This is required to express kinetic parameters of mathematical models as absolute numbers and helps to interpret fitted parameter values.

Absolute quantification was achieved by measuring the fluorescence intensity of single RNA molecules. When MCF7–PCP_GREB1_ex2_c16_Cre cells were imaged at maximum excitation energies, low intensity spots were visible in the vicinity of transcription sites (Figure 16A, left), which represent finished transcripts diffusing in the nucleoplasm. Because these spots were not visible at low excitation energies, I determined the intensity of these spots at high excitation and scaled the derived mean intensity of a single RNA to a value that would be expected at live-cell imaging conditions at low excitation (Figure 16A, right). I derived a value of either 23.3 or 32.2 as fluorescence intensity for a single transcript, depending on the imaging conditions, and used these values to calibrate fluorescence intensities of transcription sites.

As an independent method for absolute quantification of single RNA intensities I matched the intensity distributions of transcription sites from live-cell imaging at 100 pM E_2 with the distribution of spot intensities from smRNA FISH at the same E_2 concentration. Individual transcripts were visible in smRNA FISH images with excellent signal-to-noise ratio, which allowed for reliable quantification (Figure 16B, left). These intensities were used to infer the absolute number of nascent transcripts at the transcription sites. Because the intensity distributions of transcription sites should be comparable when measured at the same E_2

concentration, I matched the distributions from both experimental methods and derived an intensity of a single transcript under live-cell conditions (Figure 16B, right). Through matching of the cumulative intensity distributions, I derived a value of 17.3 a.u. as fluorescence intensity for a single RNA. This value is very close to the value of 23.3 that was derived from live-cell imaging above and confirmed that the measurement of individual PP7-containing transcripts at high excitation energies in living cells delivered reliable quantification.

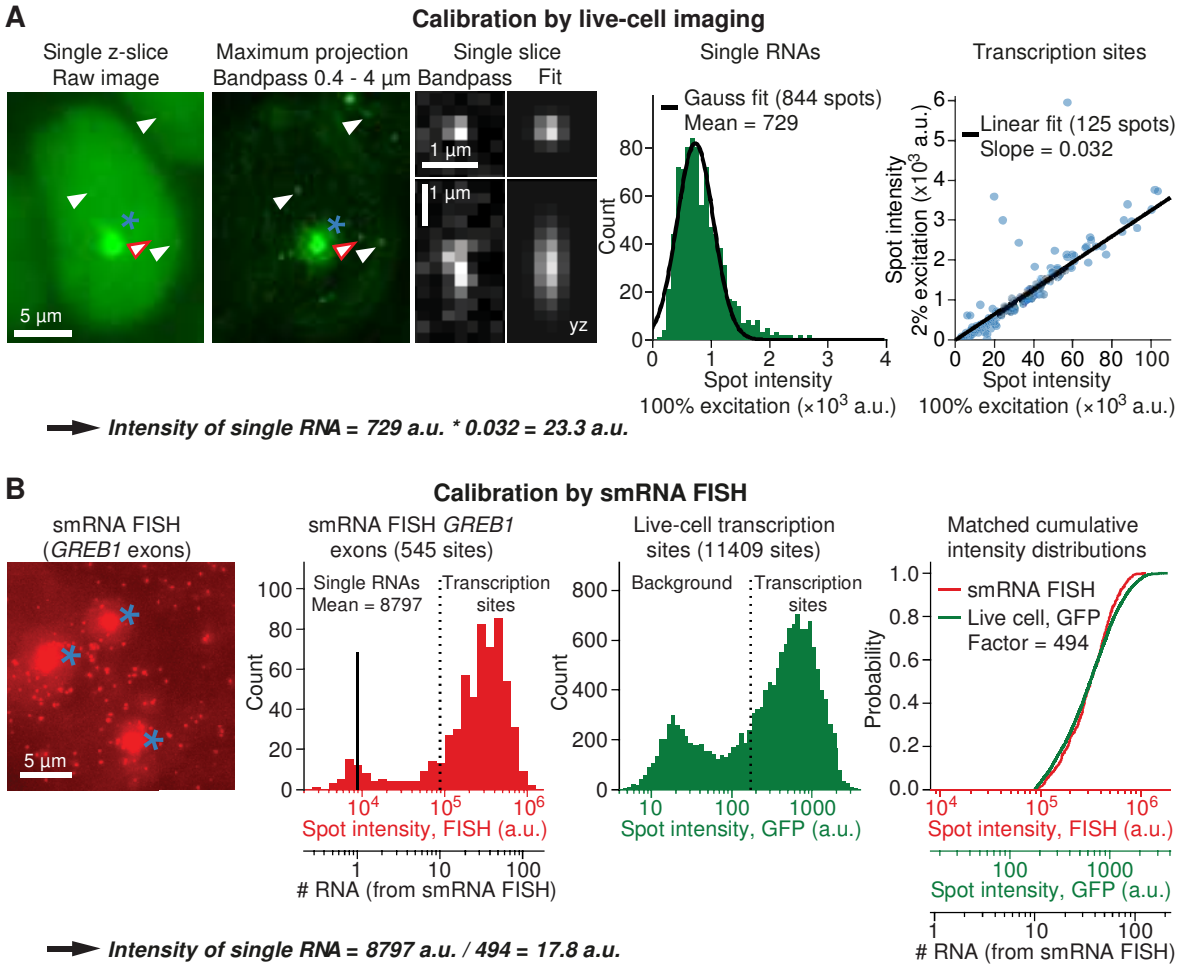

Figure 16: Calibration of fluorescence intensities. (A) Calibration by live-cell imaging of single RNA. (Left) Individual mRNAs (red arrowheads) are visible as dim spots in the nucleoplasm in the vicinity of transcription sites (blue asterisk) at 1000 pM E_2 when imaged at high excitation energies. Bandpass filtering accentuates spot-like signals. (Middle) Histogram of single RNA spot intensities. The mean of a fitted Gaussian function is indicated. (Right) Transcription sites were imaged at 2 % and 100 % excitation energy and the slope of a fitted linear function is used to relate intensity measurements between both imaging conditions. This gives rise to an intensity of 23.3 a.u. for a single transcript. **(B)** Calibration by histogram matching to smRNA FISH intensities. (Left) Image of smRNA FISH for exonic *GREB1* sequences at 100 pM E_2. Single RNAs are visible as diffraction-limited spots and transcription sites (blue asterisk) are apparent as bright foci in the nucleus. (Middle-left) Nuclear exonic foci that co-localize with GFP spots were quantified, with single RNAs representing the left peak (mean of a fitted lognormal distribution is indicated as solid black line) and transcription sites (intensity > 10 RNAs, dashed line) in the right peak. (Middle-right) Histogram of live-cell transcription site intensities at 100 pM E_2 that is used to compare intensity distributions between smRNA FISH and live-cell imaging. (Right) Matched cumulative intensity distribution of live-cell and smRNA FISH transcription sites. Scaling of live-cell intensities leads to a value of 17.8 a.u. for the intensity of a single RNA.

2.4.2 Live-cell imaging and quantification of *GREB1* transcription

Imaging of nascent RNAs is the prime tool to study dynamic regulation of transcription. It allows observation of kinetics of transcriptional bursts that are inherently stochastic and that reflect the intrinsic variability in each gene or allele. Because quantification occurs at the level of *de novo* transcription, there is no influence of RNA or protein half-lives on the observed signal. In addition, by following the activity of a single locus over time, more stable fluctuations can be studied, which represent stable chromatin configurations or the global cellular state and define gene extrinsic components of cell-to-cell variability. Discrimination between intrinsic and extrinsic noise sources is important, as it permits quantification of what proportion of the total variation is due to randomness and not under cellular control.

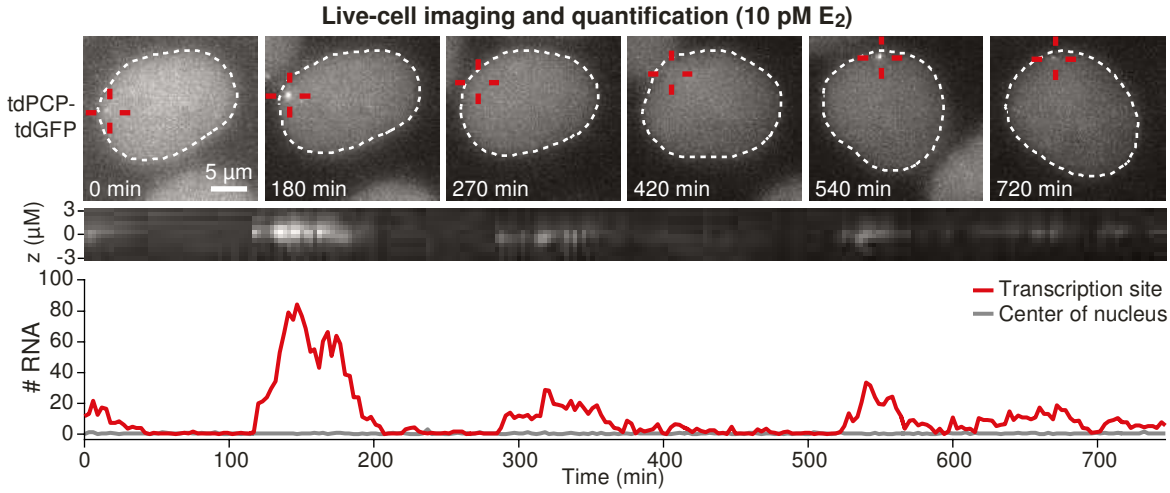

Figure 17: Live-cell imaging of transcription reveals bursts in *GREB1* RNA production. MCF7-PCP_GREB1_ex2_c16_Cre cells were imaged for 12 hours at 10 pM E_2. Transcription sites (red cross) were tracked within nuclei (dashed line). A zt-kymograph of the tracked transcription site demonstrates stable focus. The number of nascent transcripts was quantified from the images for a transcription site (red) and a control site at the center of the nucleus (grey).

I observed transcriptional dynamics of *GREB1* every three minutes over a period of 12 hours. This interval is sufficient to capture all transcriptional bursts, as it is well below the estimated ~30 minutes residence time of individual, nascent *GREB1* RNAs. At the same time, the observation time is long enough to observe multiple bursts for each cell. I tracked transcription sites within nuclei in a semi-automated way (Figure 17, top). The disappearance of foci during transcriptionally inactive periods complicated fully automated analysis and often required manual corrections. Mitotic cells and cells in which the allele replicated or moved out of focus during the movie were discarded. The fluorescence intensity within the volume at the tracked position was quantified by fitting a three-dimensional Gaussian function. As a control, a position in the center of the nucleus was quantified. The image analysis pipeline resulted in a high-quality quantification of nascent single-cell transcription (Figure 17, bottom).

The fluorescence traces showed characteristic peaks that were interspersed by periods of transcriptional inactivity in which no spot was visible. During these pauses, which occurred with an apparently random duration, the intensity was indistinguishable from that of a control site. This pattern clearly indicates that *GREB1* is produced in stochastic bursts. Further below, I analyze the duration of transcriptionally active (ON) and inactive (OFF) periods in more detail and explore possible regulatory mechanisms.

2.5 Observation of estrogen-dependent transcription

2.5.1 *GREB1* transcriptional dynamics at eight concentrations of E_2

A major aim of this study was to assess how estrogen modulates transcriptional bursts to adapt transcriptional output according to the signaling input. To this end, single-cell transcriptional profiles were recorded in more than 600 cells across eight concentrations of E_2, ranging from estrogen starvation (0 pM) to saturating induction (1000 pM) (Figure 18A). This dataset provided the basis for exploratory analysis of the ensemble of fluorescence trajectories by extracting characteristic features and global intensity profiles, analysis of the cell-to-cell variability across conditions, as well as for the derivation of a unifying mathematical model of estrogen-dependent transcription.

It is obvious from the raw datasets that transcriptional output is increased by estradiol. *GREB1* transcription sites are visible more often and show higher intensities at higher E_2 concentrations, as indicated by the heatmaps in Figure 18A. Even without E_2, a basal level of transcription is visible as short bursts in some of the cells. To further quantify estrogen-dependence, I calculated a global intensity distribution, summarizing the fluorescent intensities over all cells and time points (Figure 18B). This shows a strong bimodality in which transcription sites either are absent (spots with background intensity) or have high intensities with multiple nascent transcripts. The distribution is comparable with the results obtained from fixed cells (Figure 15) when the “background” peak is considered as no spot. In addition, the spot intensity and the fraction of time in which spots are visible (Figure 18B, right) showed an estimated EC_{50} for E_2 in the low picomolar range that is comparable to the values derived from fixed cells. This indicates excellent agreement between live-cell conditions and high-content imaging in fixed cells.

Individual cells showed transcriptional bursts throughout induction conditions with an intensity up to a value that equals 150 elongating polymerases at the gene body (Figure 18C). Bursts occur at various time intervals and the signal of consecutive bursts overlapped at high E_2 concentrations. The live-cell dataset also resolved temporally stable patterns in transcriptional activity, which were masked by individual bursts in snapshot measurements: throughout all induction conditions, a strong cell-to-cell variability (coefficient of variation between 0.36 and 0.78) is apparent in the total number of RNAs that are produced during the observation period (Figure 18A, right). Such differences might arise from stable differences in cellular state, which I will describe in detail further below.

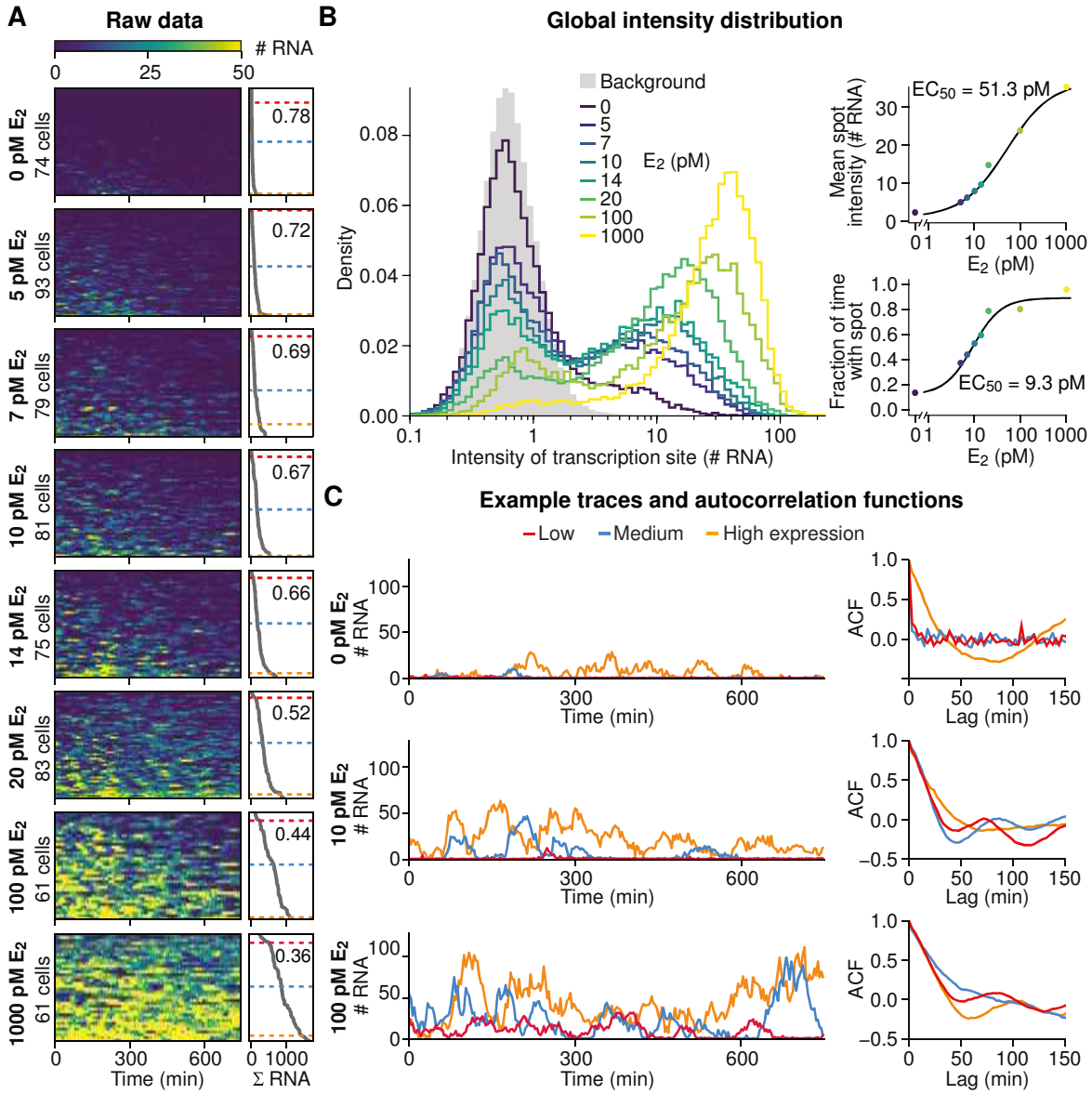

Figure 18: Dose-dependence of stochastic estrogen-dependent transcription. (A) Fluorescence intensities of transcription sites of MCF7-PCP_GREB1_ex2_c16_Cre cells are represented as heatmaps, with each row representing a single cell trajectory. Color denotes the number of nascent transcripts as indicated. For each E_2 concentration, cells were sorted from low (top) to high (bottom) total RNA output, as indicated on the right. The coefficient of variation (standard deviation/mean) is indicated in each panel. **(B)** Transcription site intensity distributions highlight the digital response of *GREB1* transcription to E_2. Histogram of all measured intensities per dataset. Mean spot intensity and spot occurrence (> 2.5 RNAs) are plotted as a function of E_2 concentration to the right. The EC_{50} of a fitted Hill equation is indicated. **(C)** Exemplary trajectories and their autocorrelation functions (ACF) for cells with high (yellow), medium (blue) and low (red) transcriptional activity for 0 pM, 10 pM, and 100 pM E_2.

In addition, at low E_2 concentrations, there was a fraction of cells that did not show a single burst throughout the 12 hours of imaging. Such non-responders are characterized by a characteristic decay at the first lag in the autocorrelation function. The autocorrelation describes how long a signal is similar to itself, hence describing a sort of “memory” in the signal. Because trajectories of non-responders consist of only memory-less technical noise from imaging, their autocorrelation decays immediately (Figure 18C top-right, red and blue curves). Based on this criterion, approximately 50 % of cells do not show transcription at 0 pM E_2. Non-responding cells were absent at E_2 concentrations above 10 pM,

indicating that all cells within the clonal population have the potential to transcribe the knock-in locus. Thus, the 10 % and 50 % of time that a transcription site is not visible at 1000 and 10 pM E_2, respectively (Figure 15 and 18B), are due to temporal fluctuations in activity in all cells rather than from individual cells that never transcribe. Responding cells show a slow decay in the autocorrelation function with a longer half-life (~ 20 min) and a shape that depends on the specific bursting behavior (Larson et al. 2011). The shape of the autocorrelation function was used during model fitting (see 2.7) as discriminator for responding cells and as a feature that describes transcriptional kinetics.

2.5.2 Extracted features describe regulation and timing of bursts

Analysis of transcription site intensities revealed that estrogen increases *GREB1* transcriptional output. So far, the time domain was only considered for autocorrelation analysis, although it contains much richer information on transcriptional kinetics. The time-resolved nature of the dataset allows extraction of information about timing, duration and intensity of each burst and to generate conclusions about regulatory mechanisms. The fluorescence trajectories are characterized by sharp increases of intensity at the beginning of each burst, during which polymerases transcribe the PP7 region. I reasoned that the slope of the curve is positive during transcriptionally active periods, even when the time in between bursts is so short that consecutive bursts overlap. Therefore, this characteristic is well suited to estimate ON- and OFF-times in transcription directly from the raw data (Figure 19A), even at high levels of induction. The amount of transcripts that are produced per burst, i.e. the burst size, was then determined from the fluorescence trace as the difference in intensity between end and start of an ON-period. The initiation rate during a burst was estimated as the ratio of burst size and ON-time.

In the context of stimulus-dependent regulation of transcriptional bursting, the number of RNAs that are produced from a gene can be either increased by increasing the number of RNAs per burst (burst size), or by decreasing the interval between bursts (burst frequency). The extracted features permit conclusions to be made whether estrogen regulates the timing, duration, or intensity of bursts, or a combination thereof. Furthermore, the shape of the distribution for ON- and OFF-times provides an estimate for how many rate-limiting steps are involved in maintaining a transcriptionally active or inactive promoter state, respectively (Suter et al. 2011, Zhang et al. 2012, Zoller et al. 2015). A single rate-limiting step would result in exponentially distributed waiting times, while more steps would produce peaked distributions (Figure 4).

I determined the duration of transcriptionally active and inactive regions and the amount of RNAs that are produced per burst for all eight datasets (Figure 19B-C). The average interval between bursts shortened from 184 minutes at 0 pM E_2 to 26 minutes at saturating E_2, indicating that the time to reactivate transcription from a silent state is regulated by E_2. This is also indicated by the average number of bursts that were observed per cell, which increased from 3 to 20. At the same time, the number of transcripts that were produced per burst increased from 5 to 15 RNAs per burst, while the ON-time slightly increased from 8 to 12 minutes. This analysis suggests that E_2 mainly controls the frequency of tran-

scriptional bursts and additionally slightly modulates burst size. Further below, mathematical modeling is used to distinguish whether E_2 indeed modulates multiple parameters of bursts, or whether a unifying model can explain estrogen-dependent transcription, even when only a single parameter is assumed to change with estrogen.

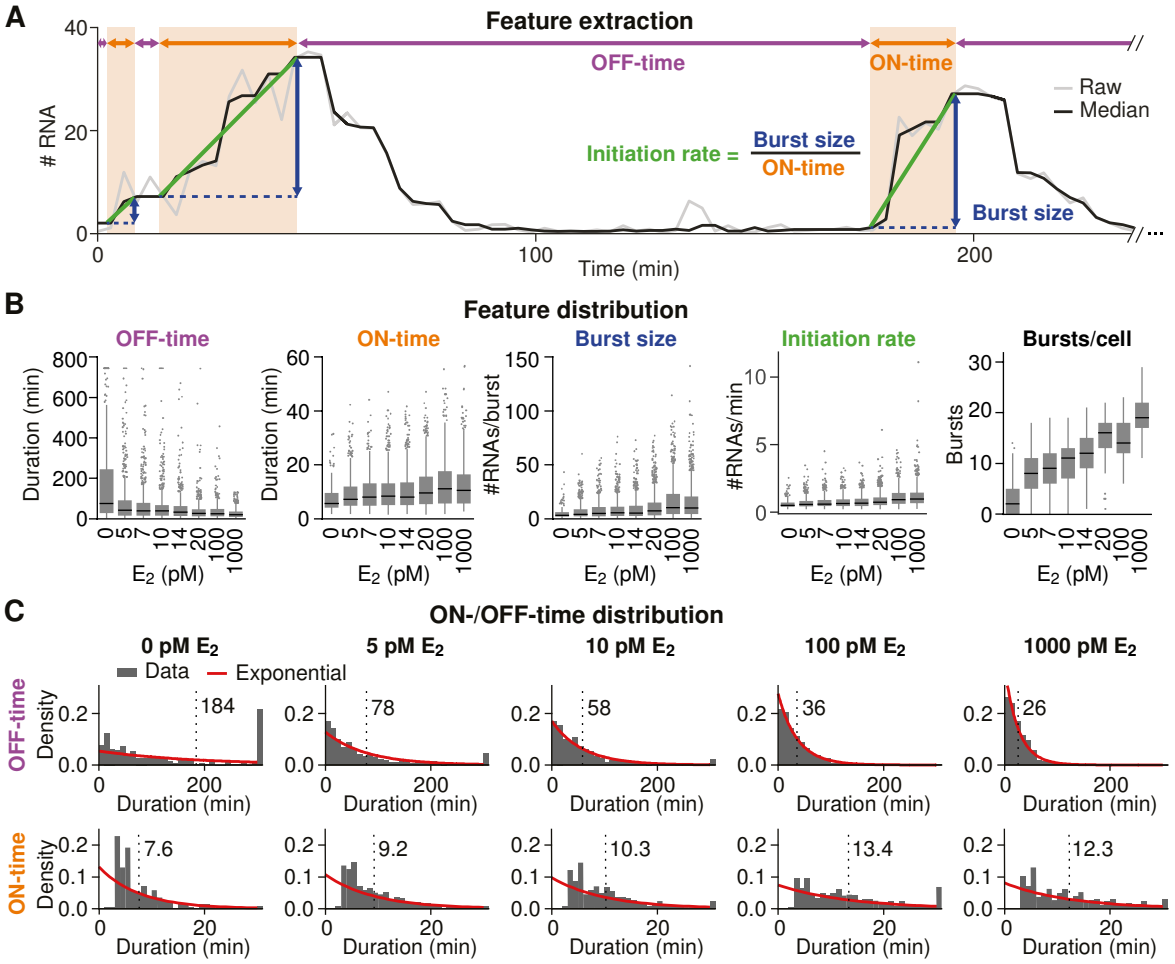

Figure 19: Estrogen-liganded ERα controls multiple features of transcriptional bursts. (A) Feature extraction from fluorescence trajectories. Fluorescence intensity trajectories (grey) were filtered by a moving median filter (black) and transcriptionally active periods were identified by thresholding the slope of the trajectory. Burst sizes are calculated as the increase in intensity during a burst and the initiation rate is calculated for each burst as the ratio of burst size and ON-time. **(B)** Estrogen-dependent changes in burst features. Features were extracted from transcription site intensity trajectories and are shown as boxplots (box = 25 % to 75 % percentile, line = median, whiskers = 1.5x interquartile range). **(C)** Distributions of OFF- and ON-times indicate single rate-limiting steps during transition between promoter states. Histograms of OFF- and ON-times are plotted along with an exponential function (red) with the same mean (dotted line, value indicated) indicating good agreement. Durations in the range of the imaging interval (3 minutes) cannot be reliably estimated leading to deviations at short ON-durations.

As individual bursts were resolved by the live-cell imaging approach, the distribution of their duration and interval is available (Figure 19C). Interestingly, the distribution of OFF-times followed an exponential function across all E_2 concentrations. The distribution of ON-times also showed an exponential decay, but deviations at short durations due to limitations in the feature extraction made this analysis less convincing. Generally, an exponential distribution for a duration of a biological event leads to the conclusion that a single rate-limiting step is involved. Such irregular timing is surprising for estrogen-dependent

transcription, given the remarkable ordered and sequential recruitment of transcription factors, co-factors, and the transcription machinery observed in ChIP experiments resulting in cyclical occupancy patterns at a target gene promoter (Métivier et al. 2003). Mathematical modeling will be used below to confirm this simple promoter model and compare its performance to models of higher complexity.

2.6 Cell-to-cell variability in *GREB1* transcriptional activity

2.6.1 Bursting characteristics vary between individual cells

All features that were extracted from raw transcription trajectories showed substantial variation for individual bursts (Figure 19B, three left panels). Most of the observed spread can be explained by the intrinsic, bursting-related, randomness within biochemical reactions leading to production of RNA. However, I also observed remarkable differences in the area under the curve (AUC) between cells within each dataset (Figure 18A and Figure 20A). This feature is a measure of the total RNA that is produced during the 12 hour imaging time-frame and encompasses temporally stable differences in transcription between cells, as fluctuations from individual bursts are averaged. High expressing cells can produce up to 10-fold more RNA than low expressing cells. For example, at 100 pM E_2, the cellular production of *GREB1* mRNA ranged from 110 to 1100 RNAs. Such variability is remarkable and cannot be explained by pure randomness through transcriptional bursting, because over such a long period with multiple bursts (on average 15 bursts per cell for 100 pM E_2) stochastic differences between bursts are averaged substantially.

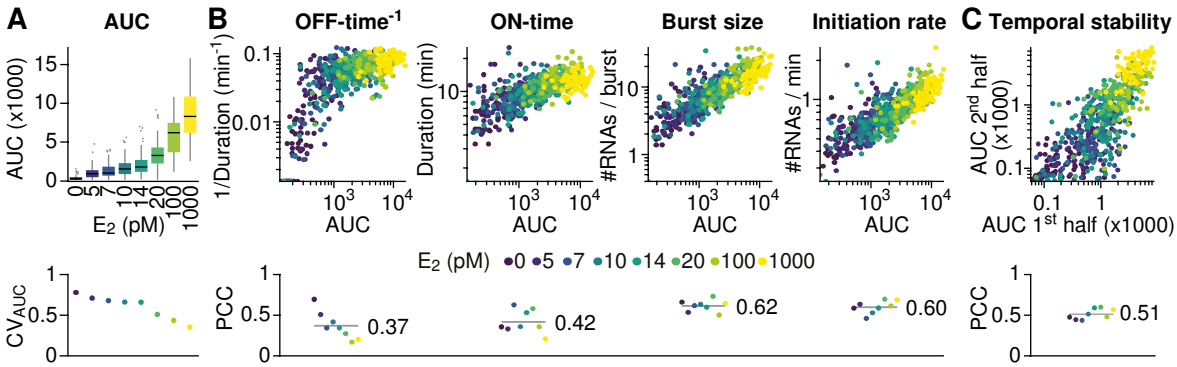

Figure 20: Variability in total RNA output is temporally stable and correlates with burst size. (A) The integrated intensity (area under the curve, AUC) was calculated per cell (box = 25 % to 75 % percentile, line = median, whiskers = 1.5x interquartile range). The CV (standard deviation/mean) for each condition is shown below, highlighting strong cell-to-cell variability **(B)** Correlations between AUC and different features are shown for all E_2 concentrations separately. The bootstrapped Pearson's correlation coefficient (PCC) is shown below as a function of E_2 with the mean indicated. **(C)** Total RNA output is stable over time. Correlation between AUC of first and second half of each trajectory (~ 6 h) are shown for all datasets with the PCC indicated below.

A more likely explanation for this heterogeneity are differences in cellular state, which lead to stable expression profiles. The state of the cell encompasses for instance the expression level of key proteins (transcription factors, polymerases, signaling factors), the signaling status (posttranslational modifications and localization of signaling factors, microenvironment), the position within the cell cycle, the metabolic state (energy supply, mitochon-

drial content), and the chromatin context at the promoter. It is difficult to assess all these parameters simultaneously and it is unknown how these non-genetic differences interplay to influence gene output. To understand how these gene-extrinsic factors modulate transcriptional bursts to achieve differences in RNA production, I analyzed the extracted features in more detail. Specifically, I determined the correlation of AUC with the mean value of features between individual cells, separately for each E_2 concentration (Figure 20B).

In the context of transcriptional bursts, the total RNA output is dependent on the number of bursts (scales with the inverse of the OFF-time) and their size (product of ON-time and initiation rate). The correlation coefficients for AUC and all extracted features in Figure 20B are positive and therefore, confirm these dependencies. The correlation with the number of bursts (proportional to the inverse of the OFF-time) decreased at higher induction. This can be explained because the number of bursts per cell is less variable when more bursts occur. This suggests that the randomness in number of observed bursts rather than cell-to-cell differences in OFF-time causes AUC correlations with the inverse OFF-time. The AUC most prominently correlated with the initiation rate and the burst size (average PCC = 0.62 and 0.60, respectively). This suggests that cellular state mainly controls transcriptional output from a burst rather than their frequency, in contrast to the regulation by E_2, which exhibits most of its control through modulation of burst frequency. Such an orthogonal control of burst size and burst frequency provides a cell with two ways to adapt transcriptional output to external signals on the one hand, and its internal state on the other hand. The dual control of transcriptional output is confirmed below through mathematical modeling.

To derive an understanding about the temporal stability of the cellular state, I determined the correlation between the AUC of the first half of each trajectory with the corresponding second half of the trajectory (Figure 20C). An average PCC of 0.51 indicates that the transcriptional output, and thus, the cellular state, is temporally stabile over the experimental timeframe of 12 hours (also compare to Figure 26C).

2.6.2 Extrinsic noise acts in *trans* to affect multiple alleles

In the previous paragraph, I concluded that cellular state alters the size of transcriptional bursts to introduce long-term correlations in transcriptional activity. However, cellular state is a complex aggregate of different variables such that the mechanism of regulation remains elusive. I wanted to distinguish whether cellular state acts locally on each allele (in *cis*, e.g. through a chromatin-dependent mechanism) or globally on many genes (in *trans*, e.g. through changes in the transcriptional or signaling machinery). To this end, I utilized the MCF7-PCP_GREB1_in2_c1 cell line that carries the PP7-sequences in two *GREB1* alleles. If both alleles within the same cell would experience the same extrinsic perturbation of bursting parameters through a *trans*-acting mechanism, extracted burst features would correlate between them. Alternatively, if both alleles would be regulated independently because of a *cis*-acting mechanism, no correlation would be observed.

I measured nascent transcription over the same 12 hour period as before, but for two transcription sites within each cell (Figure 21A). The timing of bursts was independent be-

tween both alleles, as their intensities did not correlate in time (Figure 21B). This highlights that intrinsic noise is predominant when a single time point is analyzed.

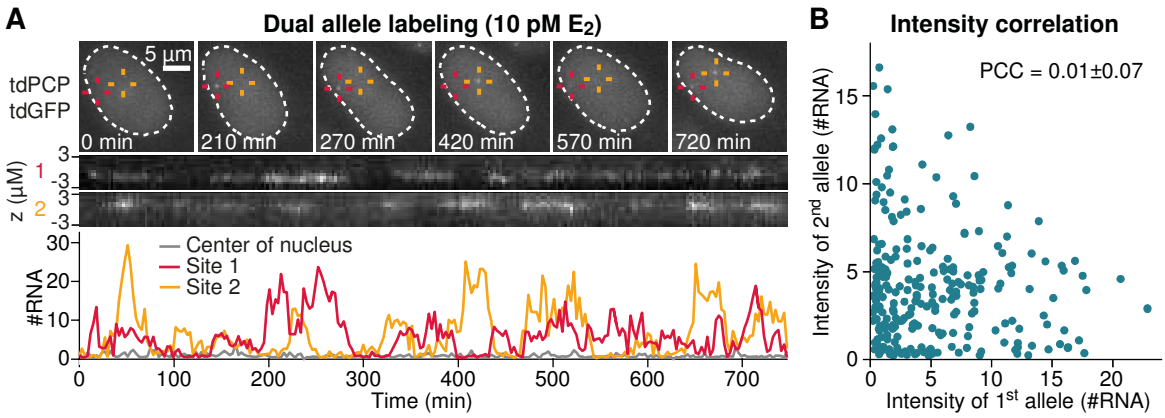

Figure 21: Transcriptional activity of two alleles within the same cell is not correlated in time. **(A)** Quantification of dual-allele transcription. MCF7-PCP_GREB1_in2_c1 cells were imaged for 12 hours at 10 pM E_2. Two transcription sites (red and yellow cross) were tracked within nuclei (dashed line). Zt-kymographs of the tracked transcription sites demonstrate stable focus. The number of nascent transcripts was quantified for both transcription sites and a control site (grey). **(B)** Transcriptional activity of sister alleles are not correlated in time. Intensities of both transcription sites from the cell in panel A were plotted, each dot representing one time point. Bootstrapped Pearson's correlation coefficient (PCC) is indicated.

I analyzed the fluorescence trajectories of two alleles within 45 individual cells (Figure 22A) and extracted features as in section 2.5.2. The mean of each feature per allele was used to calculate how correlated the features are between sister alleles. Through this time averaging, the numbers were less prone to intrinsic variability. Each of the five determined features showed positive correlations between alleles (Figure 22B). The extracted AUC, the initiation rates, and the burst sizes were highly correlated, while the duration of ON- and OFF-periods showed only weak correlations.

Correlation in the long-term productivity of each labelled allele suggests that both *GREB1* alleles are subject to extrinsic factors that act globally to control RNA output. As has been suggested by the single-allele analysis in Figure 20, these factors determine RNA output by modifying the amount of transcripts that are produced per burst. Such a global mechanism, influencing two distinct alleles at the same time, excludes a chromatin-based, local mechanism that acts to coordinately affect one or few genes. It is rather a diffusible agent that acts in *trans* to affect multiple genes by controlling how many polymerases initiate transcription from a permissible promoter state.

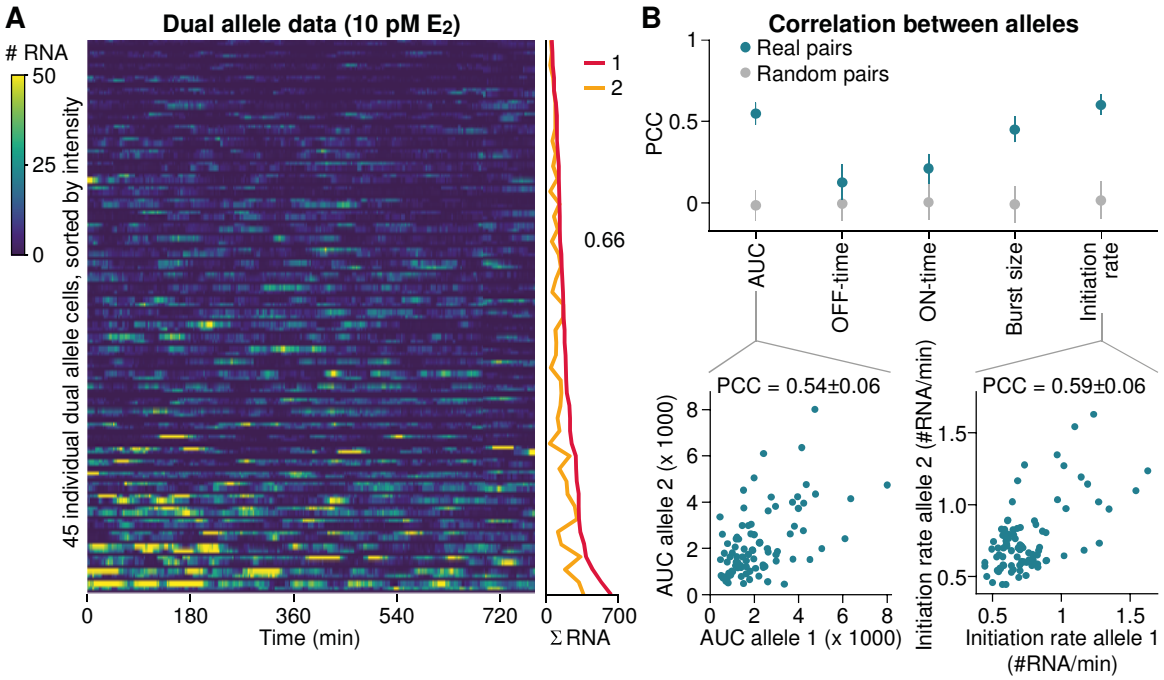

Figure 22: Cellular state controls total RNA output from *GREB1* sister alleles in *trans*. (A) Dual allele transcription in multiple cells. Transcriptional activity was quantified at 10 pM E_2 in 45 MCF7-PCP_GREB1_in2_c1 cells. The signal of two alleles from the same cell is represented as pair of rows, separated by a dark line. Cells are sorted for the total RNA output of the brighter allele from low (top) to high (bottom) as indicated on the right. The coefficient of variation across all alleles is indicated. **(B)** Total RNA output (AUC) correlates between sister alleles along with burst size and initiation rate. Pearson's correlation coefficient (PCC) between sister alleles (blue) is shown for all features with error bars denoting standard deviation from bootstrapping. Randomly reassigned sister alleles (grey) do not show correlation. Total transcriptional output and initiation rate of both alleles is plotted below as examples for strong correlation between alleles. Each allele is represented as two dots with x and y exchanged, giving rise to symmetry in the plot.

I further addressed the nature of the extrinsic noise source by following cells after cell division (Figure 23A). A diffusible factor would be distributed amongst the two daughter cells and lead to correlations in total RNA output between them. I imaged cells for 25 hours to observe cell division events and still acquire enough information from the resulting daughter cells. Interestingly, the spread in total RNA output within the first 6 hours after cell division between all cells observed at 100 pM E_2 (CV = 0.40) was comparable to the one from cells observed at 100 pM E_2 without performing cell cycle alignment (CV = 0.44, see Figure 18A). As all cells after division were in G1 phase of the cell cycle and they still displayed strong cell-to-cell variability, this indicates that variation with respect to cell cycle phases is not strongly contributing to the observed expression heterogeneity. Daughter cells showed pronounced correlation in total RNA output (PCC = 0.50, Figure 23B), which is comparable to the correlation between sister alleles (PCC = 0.54, Figure 22B) and the temporal stability within one allele (PCC = 0.51, Figure 20C). The correlation in total RNA output between recently divided daughter cells is consistent with the assumption of a diffusible factor that controls cell-specific transcriptional activity and that is partitioned during cell division. A diffusible factor could be, for example, polymerases and transcription factors available in that particular cell. It could also be related to the energy content of the cell when cells differ in mitochondrial content. It would be interesting to follow transcription in daughter cells for prolonged periods to determine how long these correlations persist.

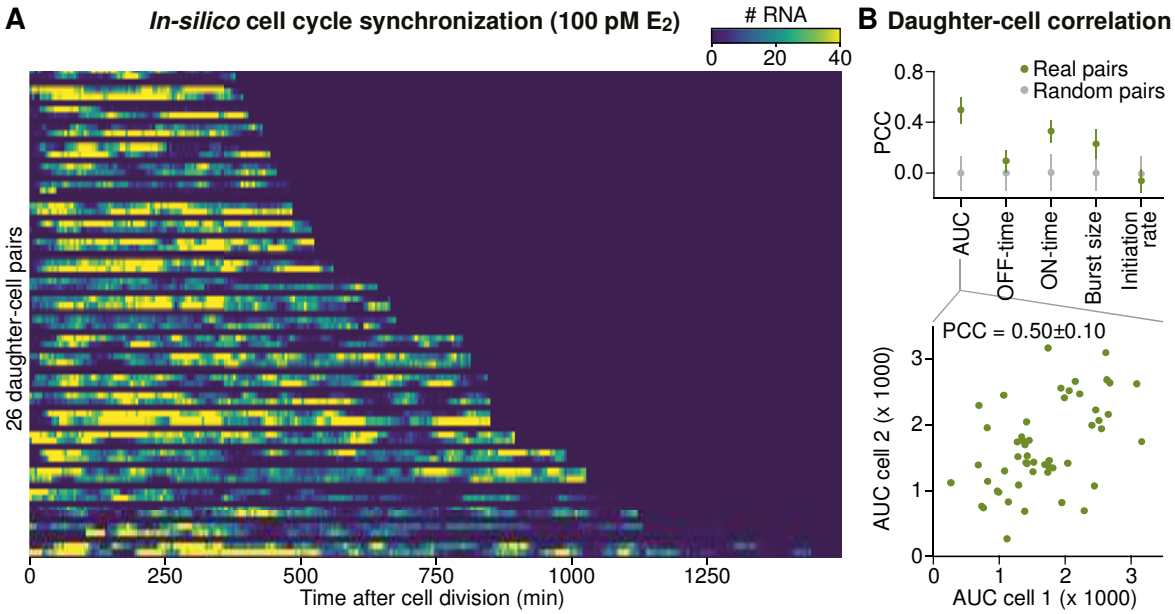

Figure 23: Total RNA output is correlated between daughter cells. (A) Fluorescence trajectories of daughter cell pairs after cell division. MCF7-PCP_GREB1_ex2_c16_Cre cells were grown at 100 pM E_2, imaged for 25 hours every 5 minutes, and daughter cells were tracked. Transcription sites were quantified and aligned to the time of cell division. The signal of two daughter cells is represented as pair of rows, separated by a dark line. Cells are sorted by their observation time. **(B)** Total RNA output correlates between daughter cells. Pearson's correlation coefficient (PCC) between daughter cells is shown for all features (green) with standard deviation from bootstrapping. Randomly reassigned daughter cells (grey) do not show correlation. Correlation of total transcriptional output is plotted below, with each cell represented as two dots (x and y exchanged), giving rise to symmetry in the plot.

2.7 Quantitative modeling of *GREB1* transcription

2.7.1 Motivation

In the previous sections, I described the generation and exploratory analysis of an extensive dataset of dynamic single-cell transcriptional profiles for the estrogen-dependent *GREB1* locus. The dataset exhibits several important characteristics:

- Stochastically timed transcriptional bursts
- Digital-to-analog transition in the global intensity histogram
- Strong cell-to-cell variability in long-term transcriptional output
- A population of non-responding cells
- Regulation of multiple burst characteristics by estradiol
- Effect of cellular state on the expression of distinct loci

The complex interplay between the individual aspects in this stochastic system can lead to non-intuitive behavior and is hard to interpret. Therefore, it is desirable to generate a model of estrogen-dependent transcription that can quantitatively describe the above-mentioned features. In the next paragraphs, I describe how mathematical models were formulated and fitted to experimental datasets to derive kinetic parameters and to distinguish competing models of promoter regulation. Furthermore, I tested predictions of the model experimentally.

2.7.2 Formulation of stochastic models with intrinsic and extrinsic variation

The probabilistic nature of transcription required the implementation of stochastic models. They had to be able to describe the temporal fluctuations in the transcriptional permissiveness of the promoter. Furthermore, these models also had to consider individual polymerase initiation events with their contribution to the observed fluorescence signal (Figure 6A). The promoter was modeled to exist in different states that are either transcriptionally active (ON) or inactive (OFF). A promoter state represents, for example, a pattern of chromatin modifications, the occupancy of regulatory proteins, or the interaction with distant regulatory elements. However, the molecular nature of the state is not explicitly considered in the model. The use of abstract ON- and OFF-states is sufficient for the occurrence of transcriptional bursts, with the most simple model consisting of a single ON- and a single OFF-state. This so-called “random telegraph” model has been used widely in the literature to explain transcriptional discontinuity (Paulsson 2005, Peccoud & Ycart 1995, Raj et al. 2006). Extensions to this model include further ON- or OFF-states that are either traversed in a linear (Neuert et al. 2013, Senecal et al. 2014) or cyclic fashion (Lemaire et al. 2006, Schwabe et al. 2012, Suter et al. 2011, Zoller et al. 2015). The multiple sequential epigenetic steps, which are reported for estrogen-dependent gene activation (Métivier et al. 2003), suggested that a cyclic model with multiple steps would be required. Therefore, five different model topologies with at least two and up to ten states were considered during model fitting (Figure 24A).

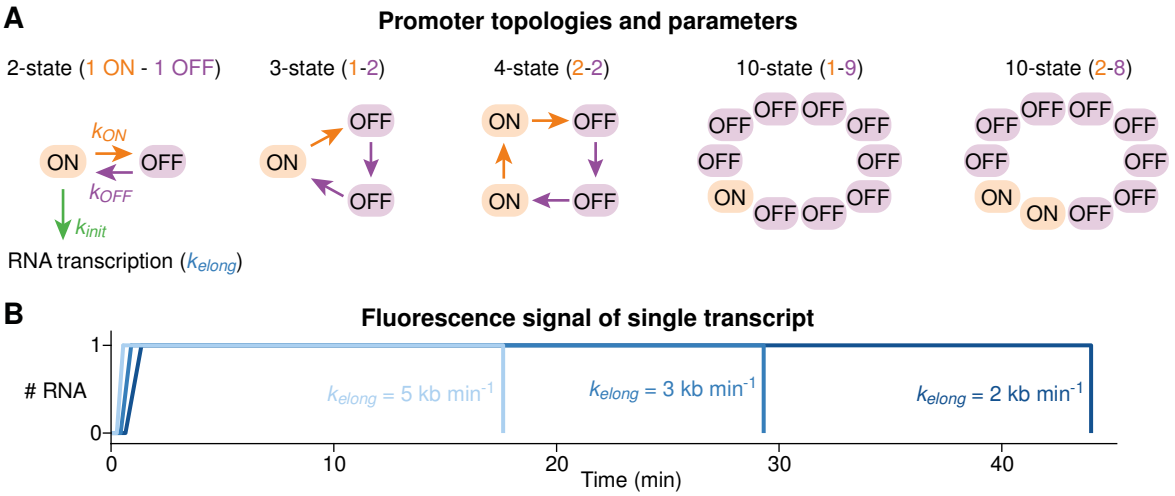

Figure 24: Topologies and kinetic parameters in models of stochastic transcription. (A) Schematic representation of cyclic promoter model topologies. Two to ten states are modeled with varying numbers of transcriptionally active (ON) and inactive (OFF) states. Transition rates between states are k_{ON} and k_{OFF} and transcription occurs from an ON-state with an initiation rate k_{init} and an elongation rate k_{elong}. **(B)** Deterministic fluorescence profile of a single transcript. Fluorescence intensity increases as the PP7 sequence are being transcribed and decreases when the polymerase reaches the end of the gene. Profiles are shown for three different elongation rates for the *GREB1* gene with PP7 sequences within exon 2.

Progression through promoter states occurs as an irreversible “ratchet-like” cycle in which each transition to the next state is characterized by the rates k_{ON} or k_{OFF}, depending on the transcriptional activity of the state. As such, the inverse of the rate is the average lifetime of the corresponding state. In case of multiple ON- or OFF-states, the individual ON- and OFF-rates were allowed to be slightly different from each other, such that variability in

rates existed, but at the same time, it was unlikely that some rates would become much faster than others, in which case bigger models would behave like smaller ones. Polymerases were assumed to initiate randomly from each ON-state with rate k_{init}. The stochastic switching between active and inactive periods is able to produce characteristic transcriptional bursts, which reflect the intrinsic variability inherent to transcription.

The temporal evolution of the promoter with all initiation events was simulated using the stochastic simulation algorithm (Gillespie 1977). For each initiation event, the trajectory of a single transcript was assumed to contribute to the fluorescence intensity in a deterministic (non-stochastic) manner that depends on gene structure and polymerase elongation (Figure 24B). Polymerase progression was modeled to be homogenous along the gene with a rate of k_{elong} that was assumed to be 3.5 kb/min, based on published polymerase kinetics (also see 2.7.7 for experimental evidence), although later also different elongation rates were allowed. The so-generated simulated trajectories of nascent mRNA numbers at the transcription site were converted to fluorescence intensities by multiplying with the fluorescence intensity of a single transcript (see 2.4.1) and adding noise of similar characteristics as the data. As a result, it is possible to generate simulated fluorescence trajectories for a given model and a set of parameters, which resemble the trajectories from the image analysis pipeline (Figure 25). This permitted comparison between simulations and experiments in a quantitative manner to fit model parameters and distinguish competing model topologies.

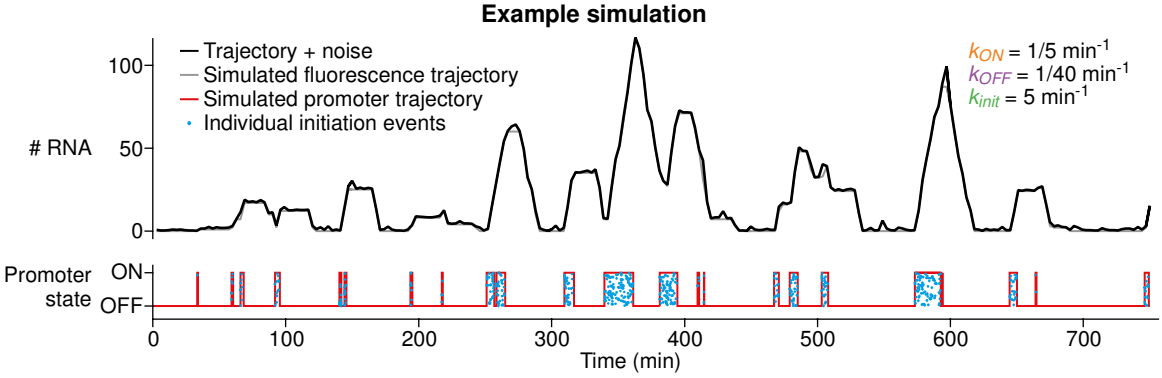

Figure 25: Stochastic simulation of transcriptional bursts. Example of a simulated fluorescence trajectory. Promoter switching (red) and initiation events (blue) were obtained by stochastic simulation (bottom). The fluorescence trajectory results from the summation of fluorescence trajectories of single transcripts (see Figure 24B) from each initiation event. The trajectory is shown before (grey) and after (black) addition of imaging noise (top). Individual bursts are observed with random timing and intensity. Parameters for simulations are indicated.

Intrinsic noise contributions alone were not able to recapitulate the above-described (see 2.6) cell-to-cell variability in total RNA output of the *GREB1* alleles in stochastic simulations (Figure 26B, grey vs. blue). Hence, extrinsic noise due to differences in cellular state had to be considered as well. Cell-to-cell variability was assumed stable over the experimental time-frame such that it could be implemented by resampling of kinetic model parameters prior to each single-cell simulation. To distinguish which model parameter had to be resampled to best recapitulate the spread in the experimental data, eight different noise sources were tested: resampling was performed for the elongation speed (k_{elong}), the

transcription initiation rate (k_{init}), the promoter ON-/OFF-rates (k_{ON}, k_{OFF}), or combinations thereof (Figure 26A). All parameters represent plausible candidates for kinetic alterations due to cellular state. The polymerase elongation rate might be dependent on the energy supply of the cell (das Neves et al. 2010) or chromatin alterations within the gene body (Jonkers & Lis 2015). The initiation rate might depend on specific enhancer-promoter contacts, promoter chromatin, or the availability of polymerases, and the switching rates might correlate with the intracellular fluctuations of transcription factor and chromatin modifying enzyme concentrations. This cell-specific resampling of parameters produced a spread of bursting characteristics in the simulated population, giving rise to a larger cell-to-cell variability in total RNA output (Figure 26B, green). Furthermore, when comparing total RNA output during the first and second half of fluorescence trajectories, simulations assuming stable extrinsic noise produced similar correlations as observed experimentally (compare Figure 26C with Figure 20C).

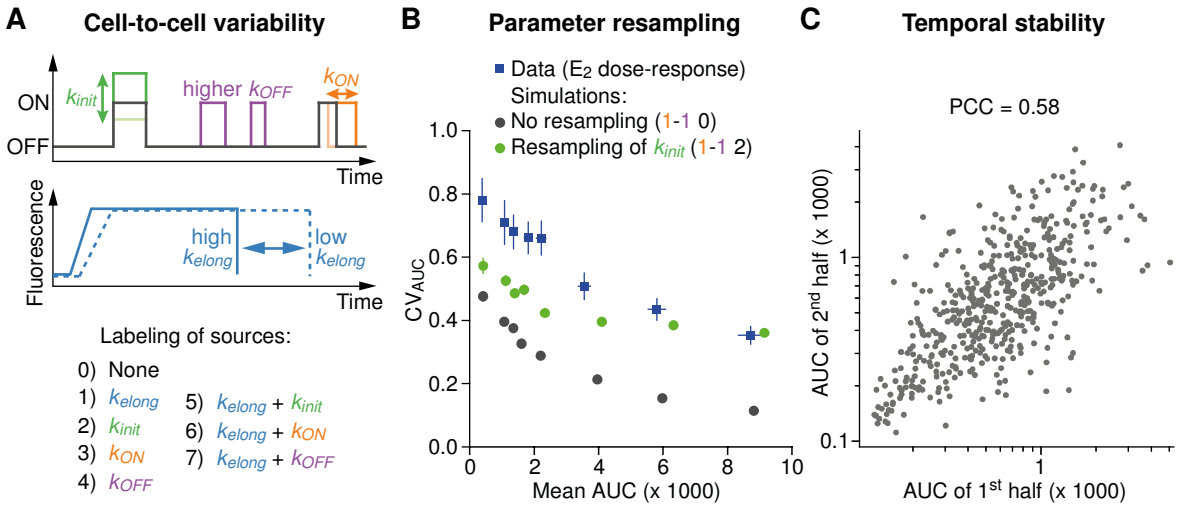

Figure 26: Implementation of cell-to-cell variability by parameter resampling. (A) Extrinsic noise sources can alter different kinetic parameters to achieve cell-to-cell variability in long-term transcriptional output. All four major model parameters individually (k_{ON}, k_{OFF}, k_{init}, and k_{elong}), as well as combinations thereof, were resampled for each cell during stochastic simulations. Models are referred to in the text and figures with the number of ON- and OFF-states, separated by a hyphen, followed by a number specifying the source of cell-to-cell variability (e.g. "1-2 3" for a three-state model with cell-specific ON-rate). **(B)** Parameter resampling during simulations better recapitulates the observed spread in RNA output. Cell-to-cell variability was calculated as coefficient of variation (CV = standard deviation/mean) for data from the E_2 dose-response (blue), and stochastic simulations without parameter resampling (grey) or with cell-specific resampling of k_{init} (green) at different OFF-times (k_{ON} = 0.33 min^{-1}; k_{OFF} = 1/500–1/7 min^{-1}; k_{init} = 10 RNAs/min). Error bars denote standard deviation from bootstrapping. **(C)** Correlation in total RNA output assuming a temporally stable extrinsic noise. Stochastic simulations were carried out (k_{ON} = 1.25 min^{-1}; k_{OFF} = 1/40 min^{-1}; k_{init} = 4 RNAs/min; model topology: "1-1 5"), and the correlation in total RNA output between first and second half of the trajectories was calculated.

The five model topologies together with the eight extrinsic noise variants lead to forty different models, which needed to be ranked in terms of plausibility during the model selection process. Model selection occurred through fitting of experimental data as outlined below.

2.7.3 Model fitting estimates model topology and parameters

Estimation of model parameters and selection between the forty different models was performed simultaneously using Approximate Bayesian Computation (ABC). ABC was implemented as a sequential Monte Carlo version (SMC ABC) (Toni & Stumpf 2009, Toni et al. 2009) in python by Stephan Baumgärtner as part of his PhD thesis (Baumgärtner 2017). He also performed all fitting of experimental data throughout this thesis. For each dataset, different combinations of parameter values and model topologies (called "particles") were generated, simulated, and compared to the experimental data, followed by iterative refinement of parameters (see 4.7.3 and Figure 27). At the beginning of the algorithm, 50.000 particles with distinct parameter combinations were generated by random sampling from the assumed prior distributions of the parameters, and stochastic simulations were carried out for each particle. A distance metric was calculated from the individual trajectories, their autocorrelation functions, and the global intensity histogram (see 4.7.3) to determine how well each particle described the experimental data. These features were chosen because they characterize the overall transcriptional output in the population, as well as the heterogeneity in transcriptional dynamics in individual cells. 2.000 particles with minimal distance to the data were used for further iterative refinement of parameters. In each iteration, particles were mutated, that is, their parameter values, including the underlying model topology, were slightly altered, new simulations were performed, and the best particles selected. The resulting final particle population yielded posterior distributions over parameters and model topologies (Figure 27, right). Because many measurement-compliant particles are present in the final population, and each of them contains all parameters, it is possible to infer correlations between them. This confines parameter values much better as compared to using distributions for each individual parameter (compare scatterplot and histograms in Figure 27, right).

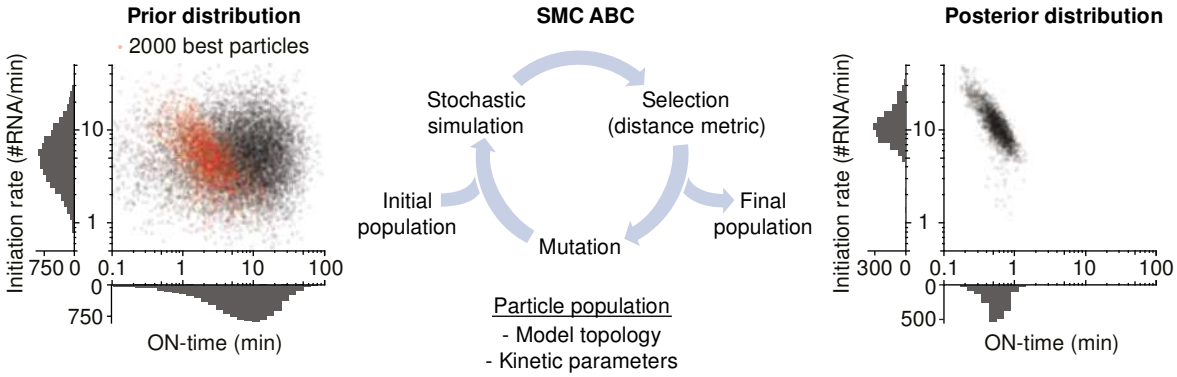

Figure 27: SMC ABC approximates parameter posterior distributions from uninformative prior. The model topology and kinetic parameters for each particle are sampled from a given prior distribution. As an example 10.000 random samples (black) of two parameters (ON-time and initiation rate) are shown for on the left with 2.000 particles (red) used as initial population for SMC ABC. This initial population is refined in rounds of simulation, selection, and mutation to yield a final population of parameters with minimal distance to the data. The posterior distribution for the two parameters is shown on the right.

The influence of pure randomness in the stochastic simulations on the distance metric was tested by repeated simulations (Figure 28A). Even with the same parameters, individual simulations differ from each other, such that the distance metric was between 0.2

and 0.8 when different simulations were compared. A value of 0.5 was chosen to represent a good fit and was hence used as a stop criterion during SMC ABC.

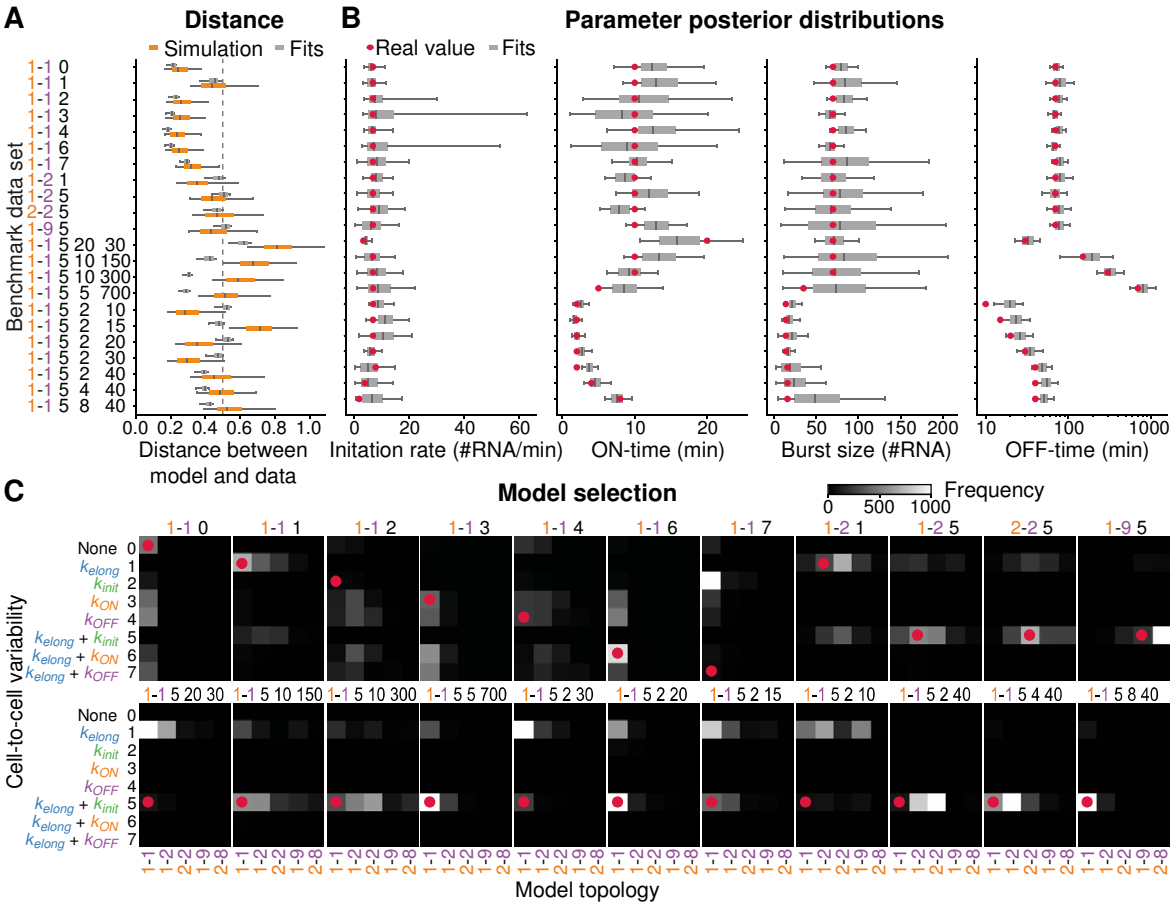

Figure 28: Benchmarking of the SMC ABC algorithm. (A) 22 sets of parameters and model topologies were repeatedly simulated and their pairwise distance was calculated to get a distribution of a best possible fit (orange). Datasets were then individually fit using SMC ABC and the distance of the best particles was calculated (grey) (box = 25 % to 75 % percentile, line = median, whiskers = 5 % to 95 % percentile). **(B)** Parameter posterior distributions were derived after SMC ABC for the main model parameters. Red dots indicate the real values used to generate the datasets. In almost all cases, the real value lies between the 5 % and 95 % percentile of the posterior distribution. **(C)** Posterior distributions over models. The count of each model in the final population of 2000 particles is represented as a heatmap with model topologies on the x-axis and extrinsic noise sources on the y-axis. Red dots indicate the ground truth.

To evaluate whether the fitting algorithm is able to distinguish competing models and finds plausible estimates for parameter values, it was tested on a set of benchmark datasets. These datasets were generated by *in silico* simulations with known parameters, model topologies, and extrinsic noise sources. This allowed an evaluation of the fitted parameters compared to known input values. All benchmark datasets were fitted using SMC ABC and the final distance between simulations and fits was close to the optimal value for each dataset (Figure 28A). Input parameters were compared with the posterior distributions (Figure 28B and C). Kinetic parameters were well estimated: In almost all cases, the real value was within the 5th and 95th percentile of the posterior distribution, and typically within the 25th and 75th percentile. Only very short OFF-times, which were below the transcript dwell-time (30 minutes), could not be reliably determined. Discrimination between model structures with different numbers of states was achieved, although closely related topolo-

gies, for example a two-state and a three-state model, were not well discriminated. Selection between extrinsic noise sources was also possible, however, proved difficult when OFF-times were short, and sometimes a single noise source was preferred when a combination was the ground truth. As such, the SMC ABC algorithm provides a useful tool to estimate parameter values and to discriminate model topologies. Hence, it was used to fit the acquired datasets.

2.7.4 Parameter fitting quantifies dose-dependence of burst kinetics

All eight datasets from different E_2 concentrations were fitted individually, that is, different parameter combinations were allowed for each condition. Features of the best particles of the final population were compared to the real datasets (Figure 29A). SMC ABC was able to approximate the features of the data, with small deviations in the half-live of the autocorrelation function. How well the fits represent the data is also reflected in the distance measure (Figure 29B), which is close to the best possible value of ~ 0.5 (compare Figure 28A). Only the two datasets at the highest concentrations of E_2 could not be fitted similarly well. The high value of the distance measure in these two datasets is caused by a poor approximation of the autocorrelation function, probably because the nascent RNA signal of consecutive bursts overlaps at low OFF-times.

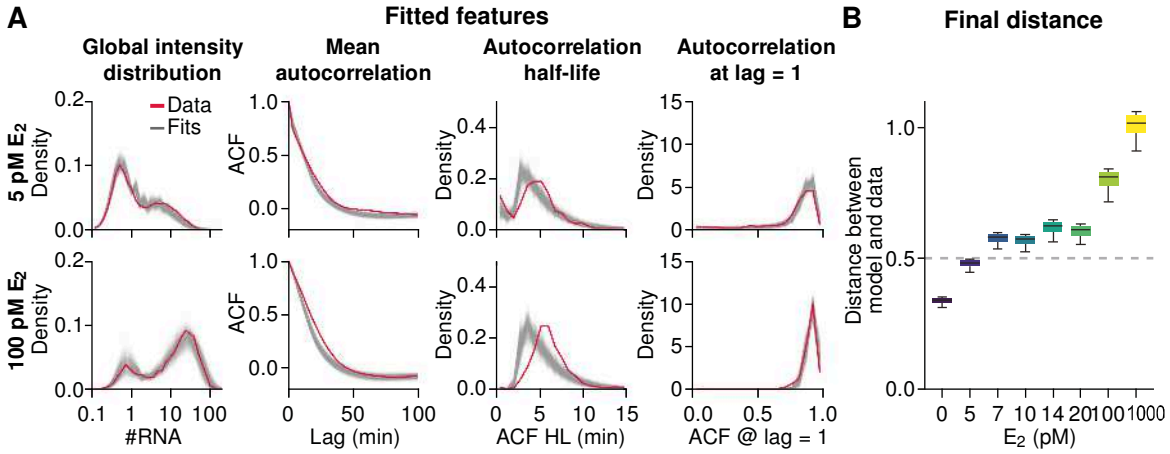

Figure 29: SMC ABC can fit major features of the data. (A) Features of the best 500 particles from fits to the 5 pM E_2 and 100 pM E_2 datasets (grey) and the data (red). All features are well approximated by the fits with slight deviations to shorter autocorrelation half-lives. **(B)** Distance measure of final particles for all datasets.

The fact that important dynamic features of the experimental data were well approximated with the proposed models of promoter progression and extrinsic noise was reassuring and suggested that the fitted parameter values indeed quantify burst kinetics. The resulting parameter posterior distributions therefore provide insight into estrogen-dependent transcriptional regulation (Figure 30). The posterior distributions for the OFF-time displayed a clear trend to lower durations at higher E_2 concentrations (Figure 30A). At 0 pM E_2 the estimated mean OFF-time was 300 minutes, while at saturating induction the value dropped to 10–20 minutes. This result highlights that E_2 regulates transcriptional output by modulating the frequency of transcriptional bursts. In addition to the OFF-time, the burst size

showed an increase with E_2 from about two RNAs/burst to ten RNAs/burst. The ON-time was slightly below one minute throughout all datasets and therefore, does not seem to be regulated by E_2. The regulation of OFF-times and burst size is in agreement with the results that were obtained by extracting features from the raw datasets (compare 2.5.2), though the absolute values differ slightly.

In addition to inference of kinetic parameters, the advantage of the mathematical modeling approach and the SMC ABC fitting procedure is that it also provides information about the underlying model structure. The posterior distribution over all 40 different models displayed a strong preference for small models throughout the datasets. Almost 70 % of all particles contained a two-state model (topology "1-1") and 85 % at most three states ("1-1" or "1-2"). A two-state model also agrees well with the observation that OFF-times follow an exponential distribution (Figure 19C). Model selection further allowed inference of how cellular state affects transcription. As sources of extrinsic noise, the model selection chose a combination of transcription elongation kinetics (k_{elong}) and initiation rate (k_{init}), denoted as model variant "5". This extrinsic noise source was the only one that appeared consistently in all datasets (Figure 30B). The combination of this extrinsic noise variant with the two-state model ("1-1 5") was present in 42 % of the particles when summarizing over all E_2 concentrations. Thus, the "1-1 5" model clearly outcompeted the "1-2 5" and "2-2 5" topologies (10 and 3 %, respectively). In conclusion, model selection clearly favored the "1-1 5" topology.

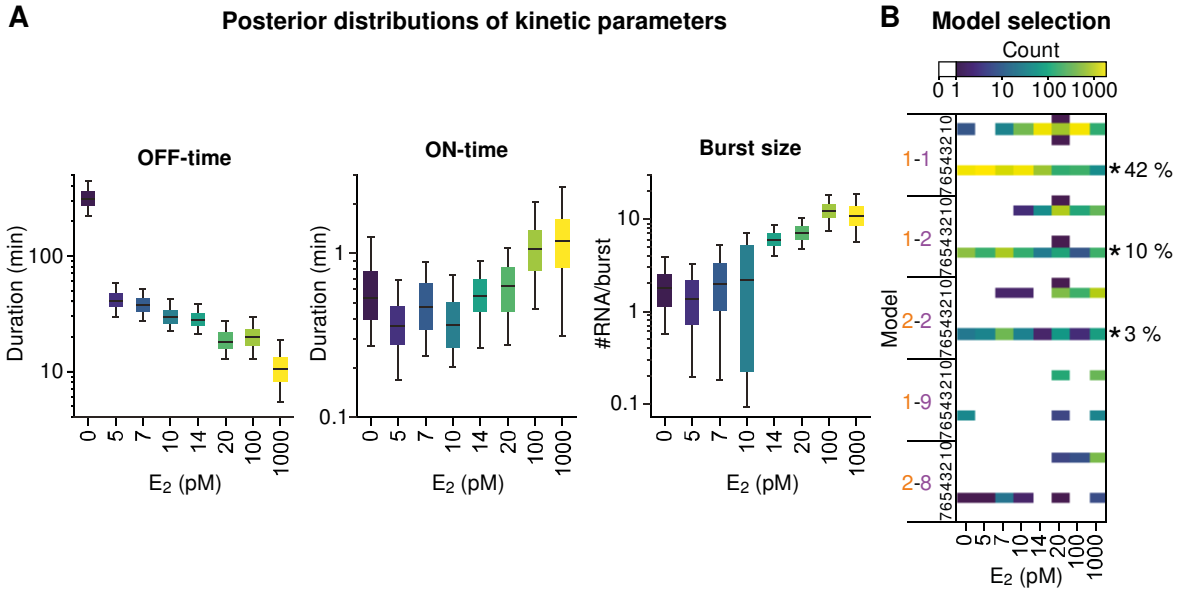

Figure 30: Marginal parameter and model posterior distributions of individual fits. (A) Posterior distributions of kinetic parameters after SMC ABC for all eight datasets (box = 25 % to 75 % percentile, line = median, whiskers = 5 % to 95 % percentile). **(B)** Frequency of all forty models in the posterior distributions. Only three models (asterisk) were found in all eight posterior distributions with the "1-1 5" model being chosen most often.

According to this result, the cellular state alters two parameters of transcriptional bursts irrespective of stimulus conditions. In line with this result, we found that the initiation rate correlated well with total RNA output, when extracted directly from the data (Figure 20) and furthermore, correlated between two alleles in the same cell (Figure 22). The selec-

tion of this parameter in explaining cell-to-cell variability is therefore plausible. Interestingly, assuming differences in initiation rate alone could not explain the data well enough. Only a combination with cell-to-cell variability in the elongation kinetics provided a good fit. This feature was not available from direct extraction and suggests that the post-initiation kinetics during transcription also vary between cells.

2.7.5 Extrinsic noise is recapitulated in simulations

I tested the extrinsic noise prediction of the model using dual allele labeling, which characterizes the correlated effect of the cellular state on gene expression. To this end, experimental data from the dual allele cell line (Figure 22) was compared to dual allele simulations, which were carried out with parameter values obtained from the posterior distributions of the 10 pM E_2 dataset. Specifically, two stochastic simulations were performed for each sampled cell-specific combination of initiation and elongation rate, because this accounts for the effect of a *trans*-acting extrinsic noise source that coordinately affects both sister alleles.

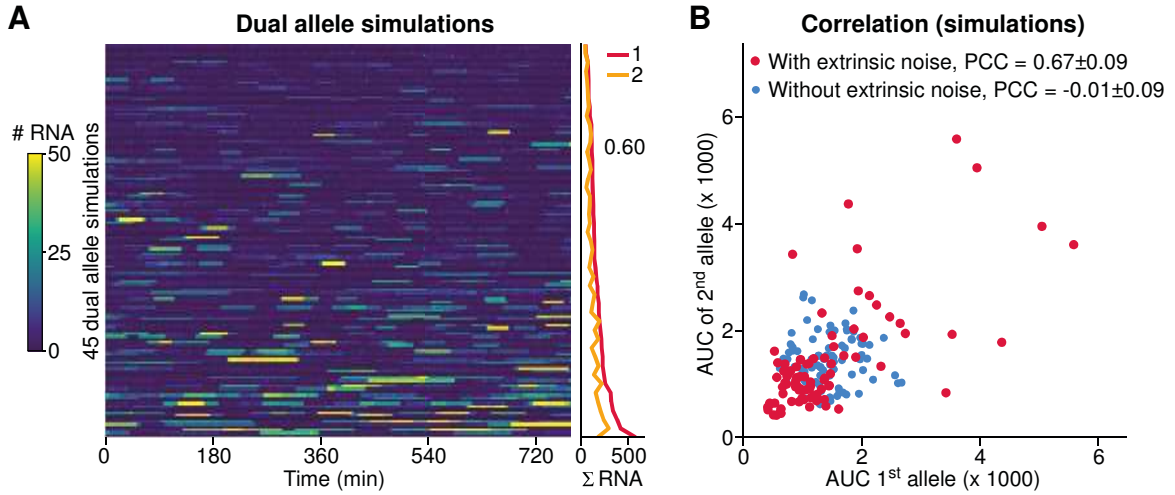

Figure 31: Simulations quantitatively confirm the contribution of extrinsic noise to correlations in total RNA output between sister alleles in the same cell. (A) Dual allele simulations. Sister alleles were simulated using the same extrinsic noise realization for both alleles with parameters from the model fit to the 10 pM E_2 dataset (Figure 30) (k_{ON} = 1.67 min^{-1}; k_{OFF} = 1/50 min^{-1}; k_{init} = 16.7 RNAs/min; model topology: "1-1 5"). Each sister allele pair is represented as pair of rows, separated by a dark line. Cells are sorted for the total RNA output of the brighter allele from low (top) to high (bottom) as indicated on the right. The CV across all alleles is shown. **(B)** Simulations incorporating extrinsic noise recapitulate correlation in transcriptional output between alleles. Dual allele simulations were performed as in panel A, either with (model topology: "1-1 5", red) or without (model topology: "1-1 0", blue) resampling of transcription initiation and elongation rates between cells. The correlation between the area under the curve (AUC) of both simulated alleles was calculated as Pearson correlation coefficients (PCC) and is reported as mean ± standard deviation from bootstrapping.

The simulated trajectories (Figure 31A) resemble the experiments, with both alleles showing uncorrelated bursting, and with correlations in total RNA output (PCC = 0.67, Figure 31B, red) quantitatively matching the experiment (PCC = 0.54, Figure 22B). Dual allele simulations without parameter resampling did not yield a correlation in total RNA output (PCC = -0.01, Figure 31B, blue). This suggests that parameter resampling achieves realistic correlations in total RNA output and that SMC ABC chose plausible parameter perturbations for the modification of transcriptional output by the cellular state.

2.7.6 A unifying model of estrogen-dependent transcription

It is likely that estrogen modulates specific steps in transcriptional activation. Therefore, I asked whether a common model topology and parameter set could explain the data from all eight E_2 concentrations, while assuming that only a single parameter changes with E_2. The model selection (Figure 30B) revealed that small promoter models with variation in k_{init} and k_{elong} explain the observed intrinsic noise due to transcriptional bursting, as well as the cell-to-cell variability in transcriptional output. This suggests that a common model topology exists. Because the “1-1 5” model variant was represented most frequently in the model selection and contains the smallest promoter topology with fewest parameters, it was decided to use the “1-1 5” model as the common model topology for all datasets. Fitting was then performed for all eight datasets simultaneously (“global fit”) to estimate kinetic parameters. To examine whether the complete range of induction could be described by a change in only a single parameter, fitting was performed with kinetic parameters being the same for all datasets, except for either k_{init} or k_{OFF}, which were fitted to each dataset individually (see 4.7.5 for details). This represents burst size or burst frequency modulation, respectively, the two modes of transcriptional regulation in a bursting context.

The global fitting approach for the per-dataset varying OFF-time yielded a distance metric that is only slightly worse than the sum of distances of the individual fits from 2.7.4 (Figure 32A). Considering the reduction of free parameters from ≥ 32 to 11, it is suggested that the much simpler model should be preferred as a unifying model of estrogen-dependent transcription. Local fitting of the burst size yielded a higher distance than for the varying OFF-time, indicating that pure frequency modulation is preferred over burst size modulation as the transcriptional regulatory mechanism of ERα. It seems plausible that a transcriptional activator regulates the transition to an active promoter state rather than the output of a burst. This way, unnecessary switching between active and inactive states is avoided in the absence of induction. In summary, a unifying model is formulated that consists of a two-state promoter cycle in which estrogen controls the time interval between bursts and the cellular state determines polymerase initiation rate and elongation kinetics (Figure 32B).

The posterior distributions of parameters quantify the difference in OFF-time between datasets and provide estimates for the burst size and ON-time that are shared between datasets (Figure 32C). The average OFF-time was 500 minutes at 0 pM E_2 and decreased to 10 minutes at 1000 pM E_2. Throughout datasets, a burst lasted on average 0.56 minutes with 7.9 RNAs being produced per burst. These kinetic parameters are similar to the ones derived from the individual fits (Figure 30).

The advantage of the global fitting approach is that estrogen-dependent transcription can now be described by only modulating a single kinetic parameter with E_2 concentration. The fact that E_2 leads to changes in burst frequency suggests that ERα establishes a transcriptionally active promoter state, rather than controlling the duration or intensity of a burst. This result is significant as it reveals the regulatory mechanism of estrogen, and as such, provides a starting point for investigations of a molecular description of active and

inactive chromatin states. Frequency modulation has also important implications on cellular noise characteristics, which I will discuss in more detail in paragraph 2.9.

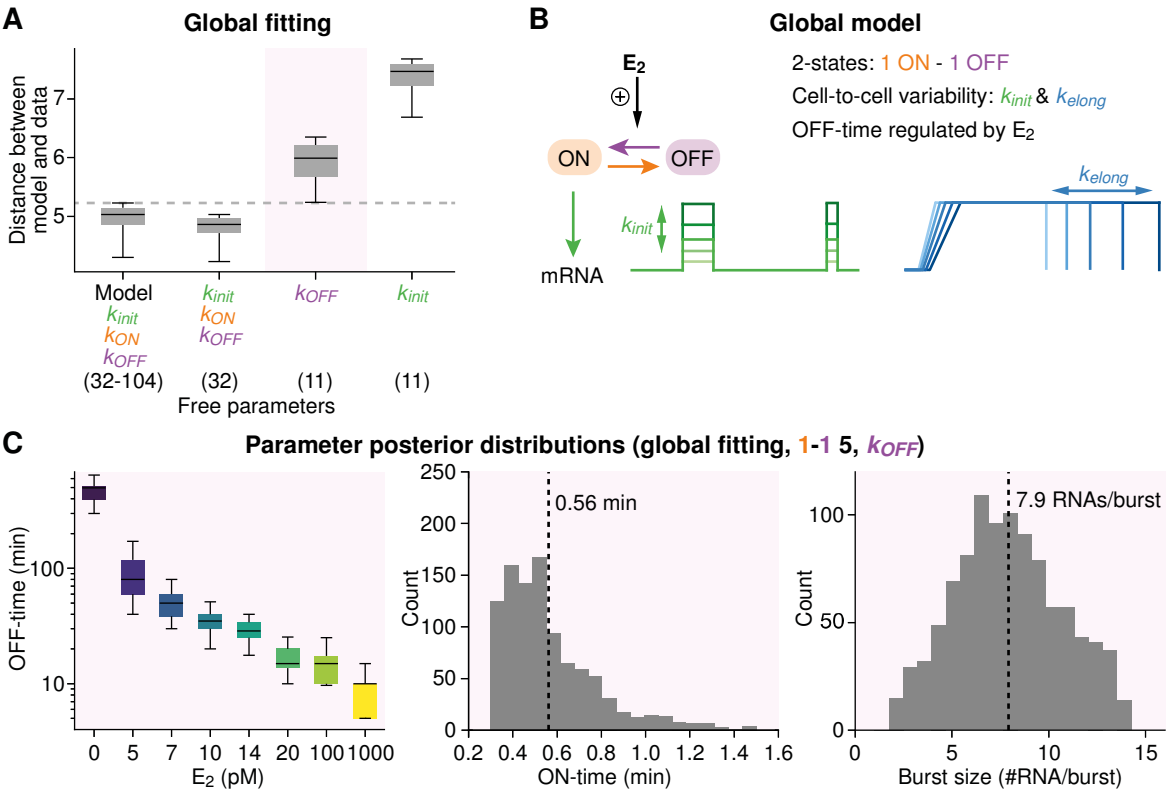

Figure 32: Global fitting reveals that only frequency is modulated by estrogen. (A) Distribution of distance measure after fitting datasets of eight E_2 concentrations simultaneously with fixed model topology ("1-1 5") and free kinetic parameters, or only allowing the parameters k_{OFF} or k_{init} to vary between datasets. The sum of distances from individual fits is shown as comparison on the left. The number of free parameters is indicated in parenthesis, with the range indicating the minimum and maximum number of parameters when different model topologies are allowed. **(B)** Unifying model of estrogen-dependent transcription. The promoter stochastically switches between transcriptionally active (ON) and inactive (OFF) states with estrogen regulating the transition into the ON-state. Individual cells vary in initiation and elongation rates. **(C)** Marginal posterior distributions of kinetic parameters from global model fitting. Distributions of OFF-time posterior distributions are represented as boxplots (box = 25 % to 75 % percentile, line = median, whiskers = 5 % and 95 % percentile). Burst size and ON-time posterior distributions are shown as histograms with the mean indicated as a dashed line.

2.7.7 Single-cell induction kinetics confirm small promoter model

The previously described unifying model of estrogen-dependent transcription makes an interesting prediction for the behavior of cells if they are synchronized in terms of transcription. When cells are grown in the absence of E_2, all *GREB1* promoters assume a transcriptionally silent state. Induction with E_2 then leads to stochastic switching to the ON-state. The unifying model consists of a single rate-limiting step in the transition to the ON-state that is regulated by E_2. Thus, the timing of the first transcriptional event is predicted to vary widely between cells. This is in contrast to a model with multiple OFF-states in which the occurrence of several reactions prior to transcription would produce synchronous timing of transcriptional events between cells, as multiple steps average the stochastic timing of individual ones. Synchronization of cells also tests the prediction that E_2 modulates burst frequencies, that is, the OFF-time. If this were indeed the case, one would

expect to see differences in the time to the first burst when cells are induced with different E_2 concentrations. Alternatively, if OFF-times are not modulated by E_2, cells should respond with similar kinetics at different E_2 concentrations and rather change the intensity of the observed signal.

Releasing cells from estrogen starvation during imaging provides an experimental approach to study induction kinetics. The single-cell nature of these experiments further allows characterization of the distributions of induction times to evaluate model predictions. Cells were grown in E_2-free conditions for three days, placed into the microscope, and after 51 minutes of imaging, transcription was induced with either 10 pM or 1000 pM E_2. Transcriptional activity was then recorded for 4 hours in about 160 cells for both conditions (Figure 33A). Almost all cells showed induction of transcriptional activity within the imaging timeframe. At induction with 1000 pM E_2, cells responded earlier as compared to 10 pM E_2. The median time to the first burst was 35 minutes at 1000 pM E_2 and increased to 78 minutes at 10 pM E_2. These results agree with predictions from the formulated unifying model in which estrogen regulates the switching speed to a transcriptionally active promoter state. However, the kinetics of induction are slower than the fitted OFF-times at the same E_2 concentrations (Figure 30). I speculate that this additional delay represents the duration of the signaling pathway, for example, the time it takes for ERα to find the promoter and to open the chromatin before the first burst takes place. Subsequent bursts might then take place with faster kinetics, though the limited observation period did not allow to sufficiently address this speculation.

The regulation of initial delay time as well as the difference in the amount of RNA produced was quantitatively recapitulated in bulk RNA measurements (Figure 33B). After starving MCF-7 cells of E_2 for three days, then inducing them with either 10 or 1000 pM E_2, samples were collected every 10 minutes for RT-qPCR. An exon-intron boundary-spanning PCR product was quantified as this measures pre-mRNA exclusively and yields a higher fold-change as compared to mature mRNA levels. The position of the PCR product in exon 2 is comparable to the position of the PP7 sequences in the exon 2 cell line that was used for imaging. The time to 3-fold induction was 52 and 29 minutes for 10 and 1000 pM E_2, respectively, and aligns well with the mean fluorescence of the PP7 signal. I also quantified the 3' end of *GREB1* (boundary of exon 32 and intron 33), to determine the delay in RNA production between early and late transcribed regions of the gene. The large genomic distance between the two locations (84 kb) allowed to estimate the rate of polymerase progression. The delay in the time of reaching 3-fold induction between the two PCR products was about 25 minutes in both conditions, which equals an average elongation rate of 3.4 kb/min. This value is comparable to other published elongation rates (wa Maina et al. 2014) and is close to the value that was assumed during modeling and stochastic simulations (3.5 kb/min). Taken together, live-imaging and RT-qPCR experiments verified that response-times depend on the estrogen stimulus and hence, confirmed frequency modulated transcription by estrogen.

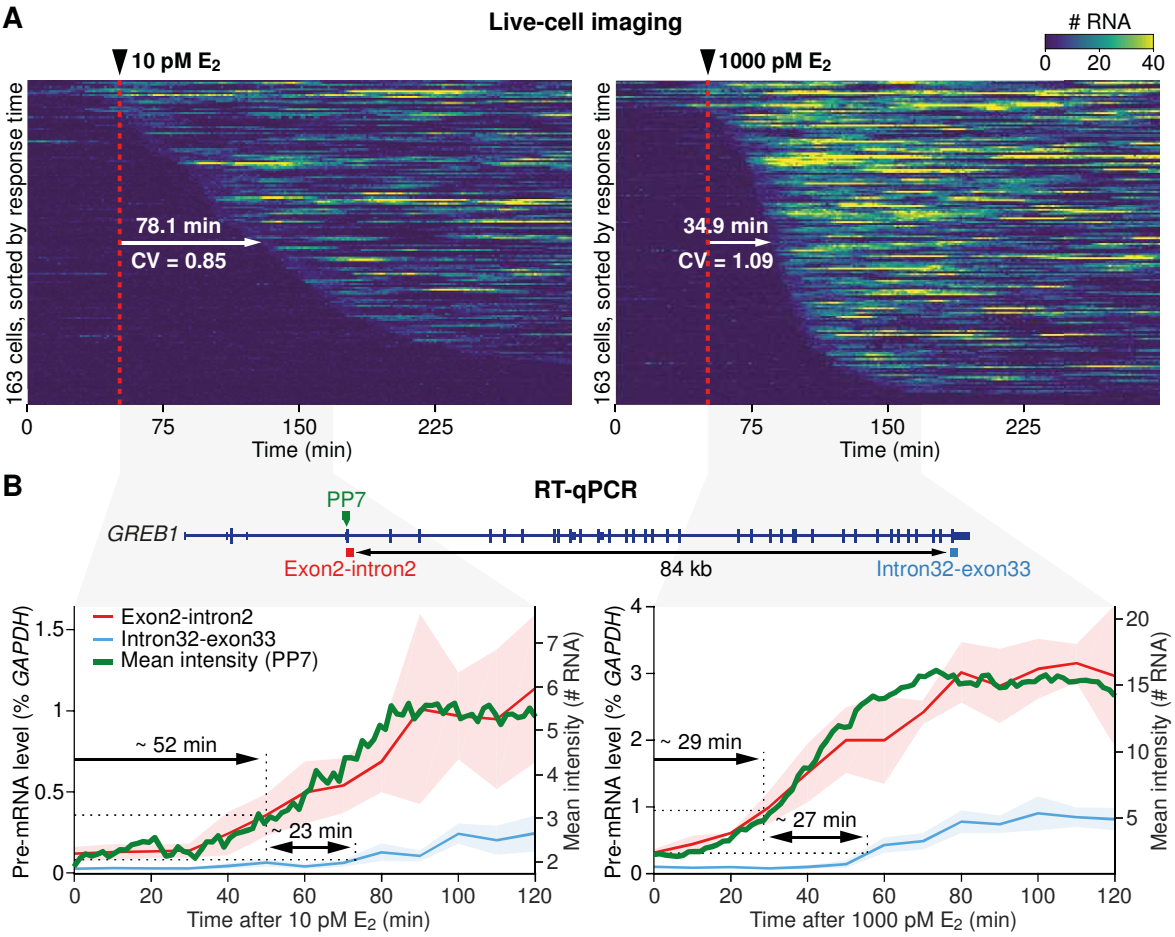

Figure 33: Synchronization experiments reveal E_2-dependent delay in transcriptional activation. **(A)** Single-cell kinetics of E_2 induction. MCF7–PCP_GREB1_ex2_c16_Cre cells were imaged at 0 pM E_2 and induced with 10 pM (left) or 1000 pM (right) E_2 at the indicated time point (arrowhead). Transcription was quantified every 1.5 minutes for 5 hours and the resulting trajectories were sorted based on the timing of the first burst. The median time to the first burst and the CV of the response-time distribution is indicated. **(B)** RT-qPCR on bulk RNA recapitulates kinetics from imaging experiments. RT-qPCR was performed for an upstream (red) and a downstream (blue) region of *GREB1* (the location of qPCR products and the PP7 knock-in is shown along the *GREB1* gene structure above). Shaded areas denote standard deviation from triplicates. The time of crossing 3-fold induction and the delay between the two PCR products is indicated.

The second model prediction was a strong variability in the time to the first burst as a result of a single rate-limiting step in the switch to an active promoter state. Indeed, individual cells differed widely in their response-time, with some cells responding within 20 minutes, while it took more than 60 minutes for others, and even longer in case of 10 pM induction. I wanted to test whether this response-time variability is better reproduced by a two-state promoter model as compared with a large (ten-state) promoter cycle. Hence, both model topologies were fitted separately to the induction datasets using SMC ABC. The fitting was performed as described earlier for the E_2 dose-response, but during stochastic simulations, all cells were modeled to initially reside in the first OFF-state and were only allowed to progress to the next promoter states after stimulation. For both datasets, the two-state model provided a better fit (Figure 34A). Furthermore, the stochastic simulations of all particles were analyzed for their response-time heterogeneity within the cell population, which was quantified as the coefficient of variation (CV, standard deviation/mean). The CV is expected to be close to one for a single rate-limiting step in gene

reactivation when the response-times follow an exponential distribution but would be lower (more synchronous timing) in larger models in which the stochastic timing of individual steps is averaged. Indeed, the experimental data and simulations of the two-state models had a CV close to one, while ten-state model simulations had lower values (Figure 34B).

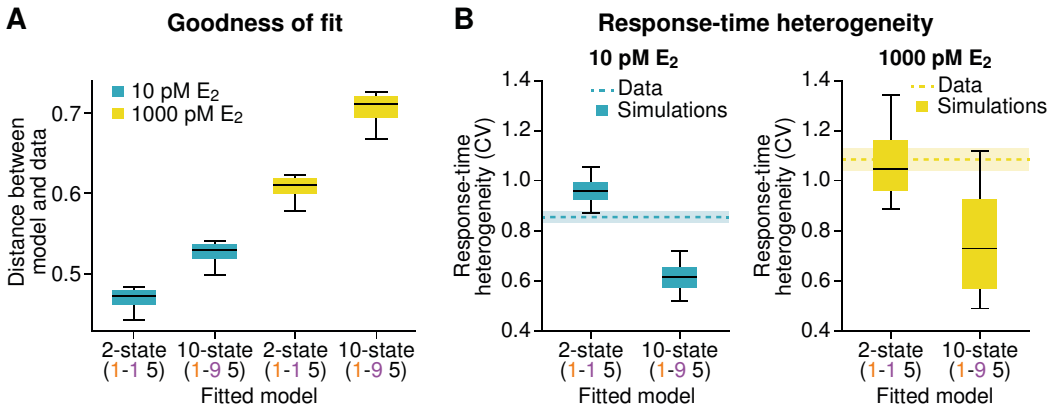

Figure 34: Small promoter models recapitulate experimental response-time heterogeneity. **(A)** Distributions of particle distances after SMC ABC on 10 pM (blue) or 1000 pM (yellow) E_2 induction datasets. The model topology was kept fixed at a two-state ("1-1 5") or ten-state ("1-9 5") model. A two-state promoter model yields a better approximation for both induction datasets. **(B)** Response-time heterogeneity in a two-state model is similar as in the data. Response-times were extracted from simulated induction experiments for all posterior particles of the SMC ABC fit. The boxplots show the distribution of coefficients of variation (CV) in the response-times over all posterior particles (central line: median, box: 25 % and 75 % percentiles, whiskers: 5 % and 95 % percentiles). Experimental CVs are indicated as dashed lines with shaded areas denoting standard deviation from bootstrapping.

This suggests, that the variability in response-times is best explained by few rate-limiting steps in gene reactivation. As noted above, the slightly slower response-time as compared to steady-state measurements argues for a delay, e.g. from signaling or opening of promoter chromatin. This, in turn, indicates further steps during reactivation, but apparently, the analysis of response-time heterogeneity is not sufficient to discriminate this. Taken together, the experiments under synchronization conditions provided another independent confirmation of the selection of small promoter models in addition to the exponentially distributed waiting times (see 2.5.2).

2.8 *GREB1* transcription requires multiple acetylation events

The unifying model for estrogen-dependent transcription consists of two promoter states, in which E_2 promotes the transition to a transcriptionally active state. However, the molecular nature of the promoter states and the reactions needed to switch between them cannot be resolved by the model, and rather necessitate perturbation experiments. I sought to understand which processes are involved in establishing a transcriptionally permissive state. To this end, I used small-molecule inhibitors to perturb the dynamics of *GREB1* transcription and determined their effect on transcriptional output. I selected inhibitors that target either estrogen signaling directly, or are associated with protein acetylation, which is relevant, for example in chromatin-mediated regulation of gene expression.

Inhibitors of estrogen signaling were chosen as controls that are known to down-regulate ERα target genes. Two anti-estrogens, ICI 182,780 and 4-Hydroxy-Tamoxifen (OHT), as well as the proteasome inhibitor MG132, were used. They either directly bind to ERα and act as an antagonist (Wakeling et al. 1991, Ward 1973), or inhibit proteasome-mediated degradation of ERα, which is known to impair estrogen signaling (Reid et al. 2003). I tested various concentrations of the inhibitors and determined the effect on *GREB1* nascent transcription by high-content imaging at a high concentration of E_2 (100 pM) (Figure 35A).

As expected, all three molecules led to decreased spot intensities and their IC_{50} was in the nanomolar range. While the two anti-estrogens almost completely abolished transcription, proteasome inhibition only reduced spot intensities by 80 % (Figure 35A, top). The inhibitory effects as such are not surprising, yet, it is interesting that the inhibition did not affect all cells similarly. Even at high concentrations of inhibitors, 20 % (ICI) or even 40 % (OHT and MG132) of cells still showed transcriptional activity (Figure 35A, bottom). This result is remarkable, as it suggests that the response to inhibitors is variegated and not all cells are targeted, which has implications on the efficacy of anti-cancer treatments. The results also highlight the power of high-content imaging in single cells to assess perturbations and heterogeneity at the level of nascent transcription. Furthermore, it motivates to test more inhibitors for their effects on E_2-dependent transcription.

With a second set of small molecule inhibitors, I assessed the importance of protein acetylation for estrogen-dependent transcription. Acetylation of histones is associated with open chromatin and active transcription (Kouzarides 2007), and these marks can serve as a specific signal for binding of regulatory proteins. Deacetylation of histones as well as acetylation of ERα itself are associated with estrogen-dependent transcription (Reid et al. 2005, Wang 2001) and hence, might influence bursting. Furthermore, recent studies showed an influence of chromatin-associated processes in regulating burst sizes and/or frequencies (Suter et al. 2011, Vinuelas et al. 2013). I chose two histone-deacetylase (HDAC) inhibitors, the carboxylate butyrate and the hydroxamic acid Trichostatin A (TSA), as well as one histone-acetyltransferase (HAT) inhibitor, C646, to assess the role of setting and erasing acetylation marks on proteins, respectively. Furthermore, I investigated whether binding to acetylated histones is important using the bromodomain inhibitor PFI-1.

In agreement with a previous study (Reid et al. 2005), treatment with HDAC inhibitors inhibited *GREB1* transcription (Figure 35B). At high concentrations of TSA or butyrate, *GREB1* transcription was essentially absent. Thus, a critical step in transcriptional activation of *GREB1* by ERα relies on deacetylation of proteins. Accordingly, the opposite effect was observed for inhibition of the HAT p300 through C646: although transcription site intensities were almost saturated at 100 pM E_2, treatment with C646 further elevated transcriptional output of *GREB1* (Figure 35B). These two observations suggest that protein acetylation negatively affects estrogen-dependent transcription. However, linking these inhibitor results to chromatin-dependent molecular processes is not straightforward because not only histones are post-translationally modified. ERα itself is acetylated by p300,

which leads to a decreased sensitivity for E_2 and increased transcription (Wang 2001). Indeed, I observed that treatment with C646 increases the sensitivity for E_2 (EC_{50} = 7 pM) when compared to the DMSO control (EC_{50} = 25 pM). Hence, the increase in transcriptional output upon C646 treatment is due to an increased sensitivity for E_2, whereas the maximal transcriptional output remains essentially unchanged (Figure 35C).

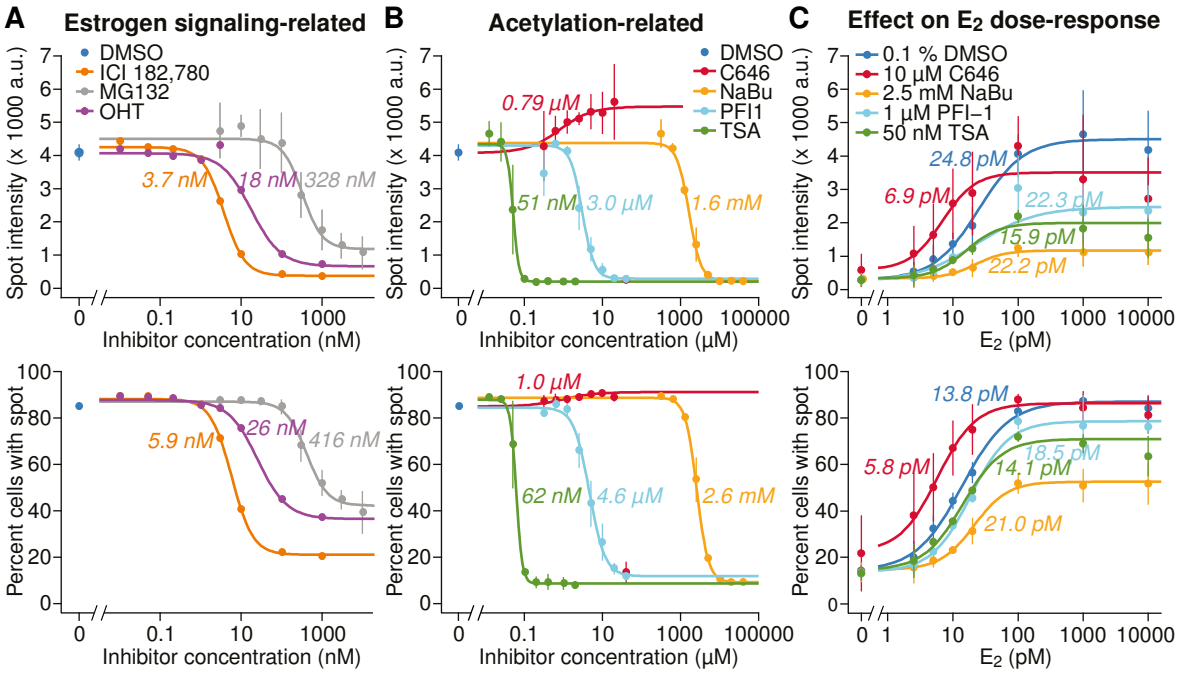

Figure 35: Effect of small-molecule inhibitors on E_2-dependent transcription. (A) MCF7–PCP_GREB1_ex2_c16_Cre cells were grown at 100 pM E_2, treated with inhibitors of estrogen signaling for four hours, and fixed prior to high-content imaging. Mean transcription site intensities (top) and percent of cells with transcription sites (bottom) are plotted, with error bars denoting standard deviation over all cells from two biological replicates (one replicate for ICI and OHT). The indicated IC_{50} was estimated by fitting a four-parameter Hill equation. All three inhibitors reduce estrogen-dependent transcription. **(B)** As in panel A but for inhibitors of protein acetylation-related processes. Inhibition of deacetylation and binding to acetylated residues inhibits transcription, while inhibition of acetylation increases transcriptional output. **(C)** Effect of acetylation-related inhibitors on E_2 dose-response. Same as in panel A but with varying E_2 concentrations at low doses of inhibitors. EC_{50} is indicated. Only C646 affects E_2-sensitivity, leading to a lower EC_{50} value.

I further assessed the role of binding to acetylated proteins by treatment with PFI-1. This molecule binds to bromodomains, and inhibits their binding to acetylated lysine residues in histones H3 and H4 (Picaud et al. 2013). Hence, it allows for conclusions on the role of acetylation "readers". PFI-1 treatment reduced the observed spot intensities dramatically and affected almost all cells, while E_2-sensitivity was unaltered (Figure 35B-C). This suggests that binding of acetylated lysine residues is as important for *GREB1* transcription as deacetylation and further highlights the role for protein acetylation in estrogen-dependent transcriptional activation.

The inhibitor datasets indicate that protein acetylation is critical for multiple molecular processes during transcriptional initiation: (I) HDAC activity controls the amplitude of the transcriptional output, with removal of an acetylated residue being important for transcription. (II) In addition, PFI-1 treatment revealed that binding of acetylated residues is likewise essential for proper RNA production. Hence, deacetylation and binding seem to affect func-

tionally opposing steps—and likely two different acetylation sites—that are important in regulating transcriptional output. Interestingly, both of these processes seem independent of the applied E_2 concentration, as the E_2-sensitivity for *GREB1* was unperturbed in both conditions (Figure 35C). (III) The acetylation of proteins (probably ERα) through p300 regulates E_2-sensitivity independently of RNA output, highlighting another acetylation step that is required in native estrogen signaling. Taken together, perturbation experiments suggest that multiple independent (de-)acetylation events determine transcriptional output from the *GREB1* locus. Such multi-step behavior was not identified through mathematical modeling and suggests that these processes are either not rate-limiting or too fast to be observable with the time-resolution and gene structure used in the experiments.

2.9 HDAC inhibition uncouples noise from mean expression

2.9.1 Expression noise and mean are inversely related during E_2 titration

Gene expression noise can be quantified by the squared coefficient of variation (CV^2 = variance/mean^2). Analytical calculations revealed that intrinsic noise scales inversely with the mean expression level for a two-state transcription model, when mean expression is only controlled by the burst frequency (Singh et al. 2010, Swain et al. 2002). The position of the resulting noise-mean trajectory is then only dependent on the burst size (dashed lines in Figure 36A).

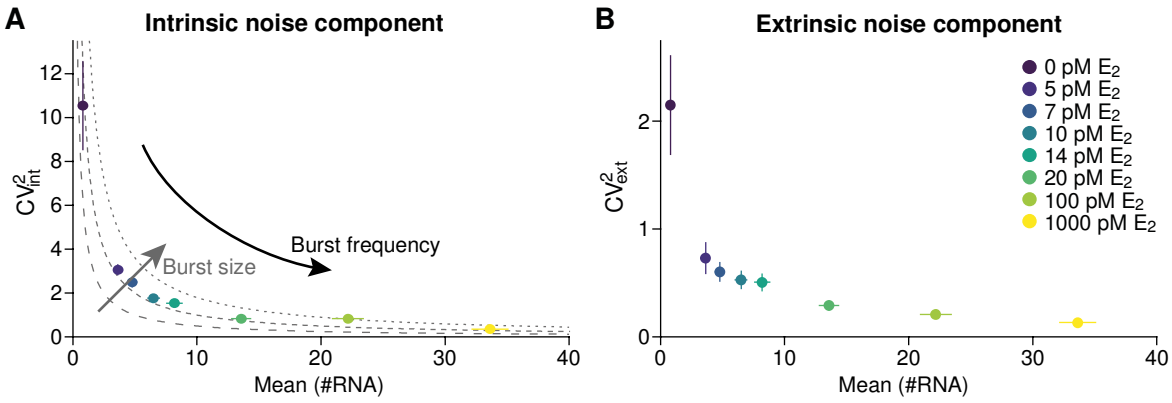

Figure 36: Decomposition of total noise reveals a dominance of intrinsic noise in promoter bursting. **(A)** Intrinsic noise scales inversely with mean transcript levels. Intrinsic noise was obtained by subtracting extrinsic from total noise (see Methods); both calculated from live-cell imaging datasets shown in Figure 18. Error bars denote standard deviation from bootstrapping. Dashed lines are inverse functions that describe the ideal behavior of pure burst frequency modulation in which the distance to the origin increases with burst size. **(B)** Extrinsic noise component is much lower than the intrinsic noise. The increase at low E_2 concentrations is likely a result of the limited observation time.

To analyze the dependence of intrinsic noise on mean expression level, the total observed noise was subdivided into contributions from intrinsic and extrinsic sources. Extrinsic noise was estimated from the CV^2 of the area under curve over all cells, as this time averaging removes most of the intrinsic noise. Intrinsic noise was then calculated as the difference of extrinsic and total noise. Intrinsic noise was much larger than extrinsic noise across all datasets (Figure 36), highlighting the strong contribution of bursts to the overall noise in nascent transcription. Furthermore, intrinsic noise was observed to be inversely

related to mean expression as it followed a trajectory of equal burst size for all E_2 concentrations (Figure 36A). This observation further, and independently, supports the frequency-modulated transcriptional regulation by estrogen.

Interestingly, also the extrinsic noise term decayed as the mean expression increased (Figure 36B). This may be caused by the limited observation period, in which long OFF-times at low E_2 concentrations result in intrinsic fluctuations that cannot be completely averaged out during estimation of extrinsic noise. Furthermore, I cannot rule out feedback mechanisms, for example when genes involved in estrogen signaling are themselves regulated by E_2 and hence, have different noise characteristics at different induction levels. Also, the cell cycle is regulated by estrogen (Prall et al. 1998), such that at low E_2 concentrations cells progress slower through the cell cycle, eventually leading to an arrest in G0 when E_2 is withdrawn. At the same time, the cell cycle affects estrogen-dependent transcription (Dalvai & Bystricky 2010). Hence, cell cycle regulation could act as another feedback mechanism, resulting in differences in extrinsic noise levels.

2.9.2 HDAC inhibition shifts noise-mean trajectory to lower noise levels

Snapshot measurements identified several acetylation-dependent events that are important for a proper transcriptional response to E_2 (see 2.8). As bursting ultimately dictates the characteristics of transcriptional noise, changes in intrinsic expression noise reflect altered bursting kinetics.

To study the effect of perturbations on expression noise, transcriptional dynamics were quantified at 20 pM E_2 with three inhibitors (butyrate, PFI-1, and C646), in concentrations that still allowed visualization of transcription sites, and with DMSO as a solvent control. The raw data (Figure 37A) revealed that, as compared with the DMSO control, transcriptional output is decreased in butyrate and PFI-1 treated cells, and increased with C646 treatment, similar to the high-content imaging experiments described in section 2.8.

I calculated intrinsic and extrinsic noise terms from these live-cell imaging experiments (Figure 37B-C). C646 treatment led to a slight increase in mean expression, while noise decreased along a curve that is theoretically expected for constant burst size. This suggests that C646 modulates burst frequency, similar to E_2, in line with the results from high-content imaging (Figure 35C), which suggested that C646 leads to an increased sensitivity for E_2. In contrast, butyrate treatment decreased mean expression levels without significantly increasing noise. PFI-1 treatment had similar effects, although less pronounced, probably because the chosen concentration was too low. Both molecules shifted the position on the noise-mean plot to curves with lower burst sizes, indicating that they primarily reduce burst size.

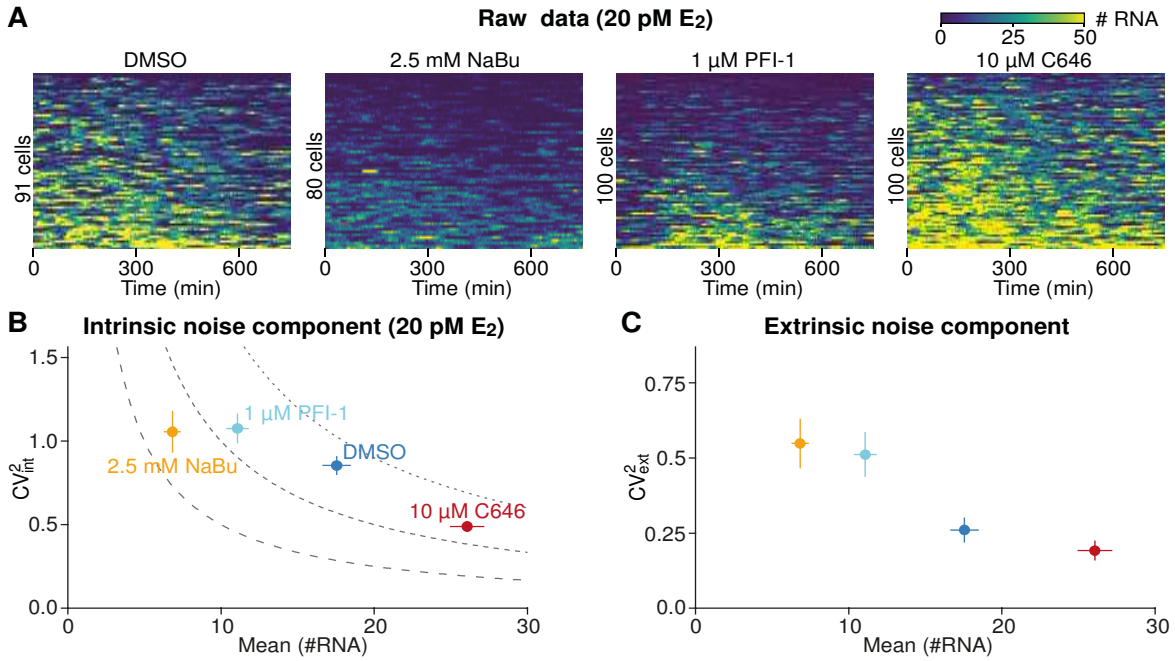

Figure 37: Inhibitor treatments affect noise-mean relation in E_2-dependent gene expression. **(A)** Fluorescence intensities of transcription sites of MCF7-PCP_GREB1_ex2_c16_Cre cells grown at 20 pM E_2 and treated with the inhibitors as indicated. Cells are sorted for the total RNA output from low (top) to high (bottom). **(B)** Intrinsic noise was quantified from the datasets in panel A. Error bars denote standard deviation from bootstrapping. Dashed lines indicate different burst sizes (same as in Figure 36A to aid comparison with E_2 titration). Low concentrations of butyrate (NaBu) and PFI-1 reduce expression with limited effects on noise. **(C)** Extrinsic noise component for experiments in panel A.

To validate the effect of inhibitor treatments on burst frequency and burst size that was apparent from noise-mean scaling, the model fitting procedure using SMC ABC was applied, separately for each condition. The posterior distributions (Figure 38A) revealed a decrease in burst size for butyrate as compared to the DMSO control, while OFF-times were only slightly changed. Similar shifts in the parameter posterior distributions were observed for PFI-1 treatment. These results confirmed that butyrate treatment reduces the size of *GREB1* transcriptional bursts, providing an explanation for the reduced intrinsic noise levels as compared to E_2 concentrations giving rise to the same mean expression. The C646 dataset in contrast, was best fit with lower OFF-times and higher burst sizes as compared to the DMSO control, supporting the assumption of modified burst frequency. As a further validation, I extracted burst features from raw trajectories as in section 2.5.2 (Figure 38B) and the trends within the shifts of the distributions were observed to agree well with the results from SMC ABC.

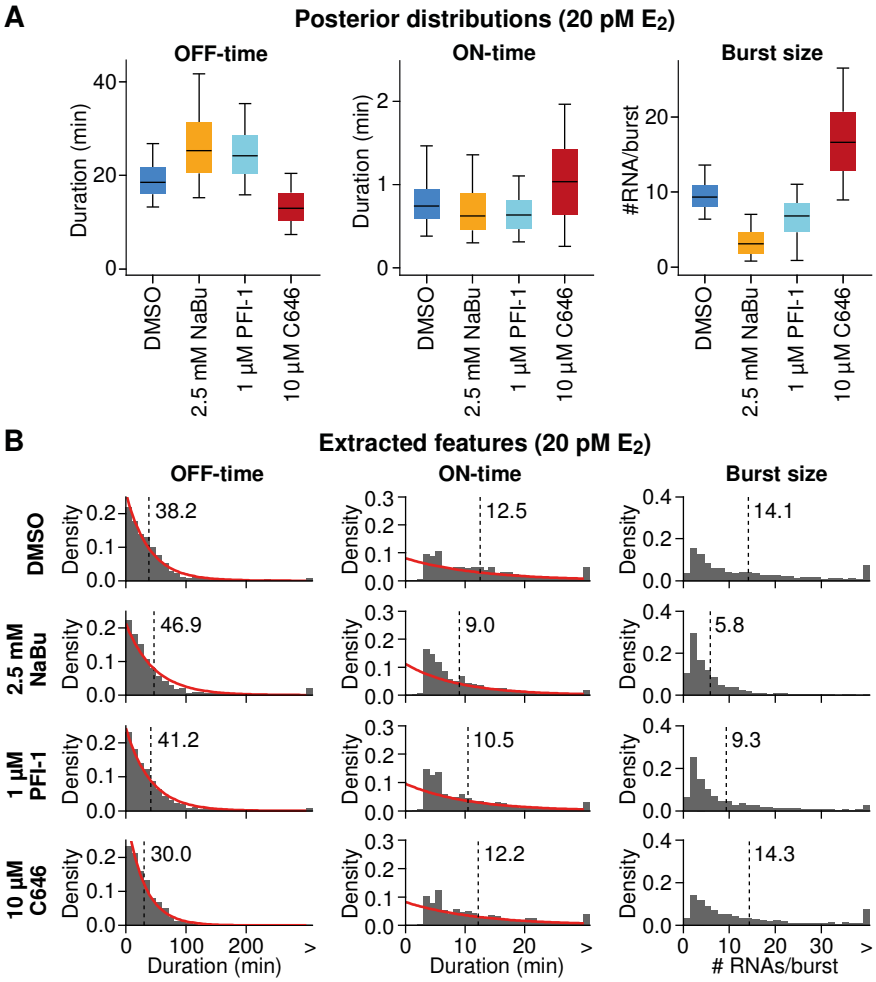

Figure 38: Small molecule inhibitors of protein acetylation alter dynamics of transcriptional bursting. **(A)** Parameter posterior distributions over all particles after SMC ABC fitting of the inhibitor datasets (box = 25 % to 75 % percentile, line = median, whiskers = 5 % to 95 % percentile). **(B)** Features were extracted from raw trajectories as in Figure 19A. The mean of the distribution is indicated (dashed lines). Exponential distributions with the same mean are shown in red.

Taken together, changes in posttranslational protein modifications influence the dynamics of transcriptional bursts and the resulting intrinsic noise. The implications of these findings are striking: the modulation of burst frequency through ERα entails a characteristic noise-mean scaling with high expression noise at low levels of expression. However, control of burst size through altered protein deacetylation or acetyl-binding can modulate this particular noise level, such that a low-noise regime can be obtained for a gene, depending on the needs of the cell. Thus, protein acetylation levels allow for fine-tuning of transcriptional output and associated noise, independent of the estrogen stimulus.

3 Discussion

Transcription is *the* major regulator for the production of cellular constituents and as such, tight regulation is necessary to ensure cellular function. However, stochastic effects and discontinuities in RNA production act to hinder consistent gene output, which results in heterogeneity between cells. It is essential to characterize and understand transcriptional regulation in such a stochastic setting, to comprehend sources and consequences of cellular heterogeneity in development and disease.

This thesis examined how cells control transcriptional bursts to adapt gene output to estrogenic signals and to describe how inherent expression noise arises and is regulated by the cell. Labeling and observation of nascent transcription of the estrogen-sensitive *GREB1* locus in living cells was used to inform mathematical models of transcription. In consequence, a unifying model was derived that describes transcriptional behavior under various experimental conditions (Figure 39). It revealed that cellular state affects gene expression on long time-scales, while intrinsic, bursting-related fluctuations dominate on short time-scales. Below, I will compare such a model with other examples of transcriptional regulation reported in the literature. I will discuss whether models of transcriptional behavior can be generalized to other genes and how the cellular state could affect expression. I will speculate on molecular mechanisms and about consequences of expression noise on cellular function and heterogeneous cancer growth.

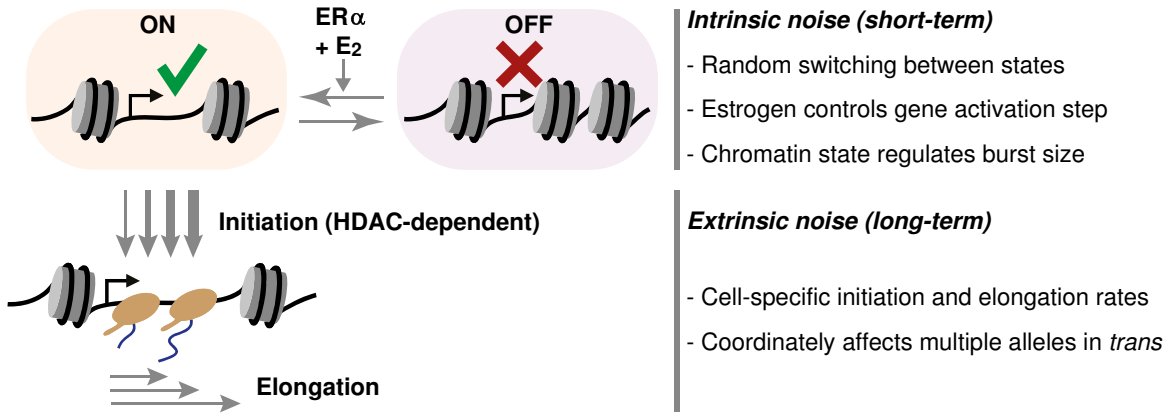

Figure 39: Regulation and noise in estrogen-dependent transcription of *GREB1*.

3.1 Two promoter states in *GREB1* transcriptional dynamics

GREB1 transcripts were observed to be produced in bursts throughout experimental conditions. Such behavior was similarly reported for other genes in various organisms (Chubb et al. 2006, Fukaya et al. 2016, Larson et al. 2011, 2013; Suter et al. 2011) and seems to be the predominant mode of transcription in human cells (Dar et al. 2012). The simplest model that allows for the occurrence of transcriptional bursts is one in which the promoter exists in two distinct states, an active and an inactive one. This so-called “random-telegraph” model (Paulsson 2005) has been used widely to model stochastic transcription and associated cell-to-cell variability (Raj et al. 2006) and was found to be sufficient for

explaining the estrogen-dependent transcription datasets in this thesis. In contrast, a simpler non-bursting, one-state model with constitutive activity, that is for example observed for housekeeping genes in budding yeast (Zenklusen et al. 2008), shows less variability in total RNA numbers per cell and is inappropriate to describe bursty genes. Extensions and variations of the two-state model exist in numerous flavors. Corrigan et al. (2016) propose a time-varying initiation rate instead of a strict separation of active and inactive periods for their observations on bursts in *actin* gene transcription in *Dictyostelium*. Their model allows for active periods with variable transcriptional output, such that it can explain transcriptional behavior in more detail, but it is also more complex, as it requires additional parameters. The two-state model that was chosen by model selection in this thesis is sufficient to explain the acquired data, but much simpler than the Corrigan model, making it more robust to parameter estimations.

Instead of varying kinetic parameters, other studies extended the random-telegraph model through incorporation of additional active or inactive promoter states. They were either added as a linear chain with reversible transitions between states (Neuert et al. 2013, Senecal et al. 2014), for example, to allow for multiple ON-states with distinct activity. Alternatively, promoter states were added to form a promoter cycle in which states are cyclically and irreversibly traversed as a ratchet-like process (Lemaire et al. 2006, Schwabe et al. 2012, Suter et al. 2011, Zoller et al. 2015). Such additional states lead to refractory periods after activation, in which the promoter cannot immediately be reactivated (Suter et al. 2011). This second extension to the random-telegraph model was used in my thesis, when cyclical promoter models with up to 10 states (Figure 24) were fitted to measured transcription data. Interestingly, the simple two-state promoter cycle best explained the observed fluorescence trajectories and intrinsic noise. A two-state model suggests that *GREB1* transcription does not have refractoriness or memory in reactivation, perhaps to allow for fast adaptation to estrogen signals.

More strikingly, large promoter cycles did not yield a good fit for *GREB1*, though they were needed to explain cyclical patterns of protein occupancy at other estrogen-sensitive promoters (Lemaire et al. 2006). In line with the preference for small promoter model topologies, extracted OFF-times were exponentially distributed (Figure 19C). Furthermore, my experiments in synchronized cells showed a rapid upregulation of transcription upon exposure to E_2, but no coherent cycles of RNA production (Figure 33). Such discrepancy could occur because single-cell transcription and biochemical ensemble experiments like ChIP (Métivier et al. 2003, Shang et al. 2000, Sun et al. 2007) measure different aspects of transcriptional regulation (i.e. promoter opening vs. nascent transcription). On the one hand, not every association of transcription factors and cofactors detectable by ChIP has to be productive and lead to RNA production. On the other hand, only a subset of cells might contribute to the observations at the population level, while single-cell experiments observe the entire heterogeneous population. Furthermore, the ordered, cyclical recruitment of cofactors were observed at the *TFF1* locus and might only be prevalent at a subset of estrogen-dependent genes, with *GREB1* not being one of them. Indeed, individual genes can differ in their promoter model topology (Zoller et al. 2015) and a single tran-

scription factor can lead to remarkably different transcriptional outcome at the single-cell level, depending on the promoter sequence (Carey et al. 2013).

Through perturbation experiments, I identified that multiple acetylation-dependent biochemical reactions are present during the ON-state, as inhibition of deacetylation and inhibition of acetyl binding reduced transcriptional outcome (Figure 35). It seems plausible that the ON-state consists of multiple steps, as the assembly of a functional transcriptional complex itself and firing of polymerases are multi-step processes. However, model fitting only identified a single rate-limiting step determining the lifetime of an active state. Apparently, all other reactions are much faster, such that they do not contribute significantly to the ON-time and were hence not identifiable. Similar conclusions hold for the transcriptionally inactive OFF-phases, which are as well characterized by a single rate-limiting step. However, they probably also contain additional, non-identifiable fast reactions, and these would be beneficial in combinatorial control of the transcription cycle (Scholes et al. 2017). Indeed, at high E_2 concentrations, when OFF-times are sufficiently short, additional states were also able to describe experimental data, as the model selection also infrequently chose bigger promoter cycles with multiple OFF-states (Figure 30B). This suggests that the rate-limiting step is no longer much slower than other events under these conditions.

3.2 Molecular nature of promoter states

A question that immediately arises from the proposed two-state model is that of the molecular nature of each state. Many studies tried to elucidate molecular determinants of bursts and revealed the importance of genomic position, binding and intracellular localization of transcription factors, polymerase re-initiation, and chromatin for burst properties in eukaryotes (reviewed in Lenstra et al. 2016). The definition of promoter states by chromatin seems tempting because unstable and reversible protein complex assembly could be translated into a metastable chromatin state (Schwabe et al. 2012). Such states could be characterized by a set of histone and DNA modifications, as well as the presence and positioning of nucleosomes. For example, repressive histone marks could lead to more compact, inaccessible chromatin at a promoter during an OFF-state. Similarly, nucleosomes could occupy transcription factor binding sites to preclude transcription factor binding. The role of chromatin in transcriptional bursting is exemplified by a study that showed no bursting (i.e. continuous transcription) when a transgene was expressed from a plasmid but bursty expression after integration into the genome (Larson et al. 2013). Furthermore, chromatin associated processes are critically involved in dynamic estrogen-dependent transcription (Kangaspeska et al. 2008, Métivier et al. 2003, 2008; Shang et al. 2000, Sun et al. 2007).

Through perturbation experiments, I identified a role for protein deacetylation during transcriptionally permissive phases (Figure 35). In addition, the active state is characterized by an acetylated residue, which has to be bound via bromodomains to allow for transcriptional activity. As such, I unambiguously identified a role for protein acetylation in distinguishing transcriptional permissiveness and inactivity of a promoter. However, the multi-

tude of histone and non-histone targets of acetylation do not allow conclusions to be made about specific protein targets. Other studies also used perturbation of histone acetylation and found that bursting kinetics are affected, though with different outcome depending on the gene and experimental system. For instance, Suter et al. (2011) report longer duration of OFF-periods after TSA treatment at a prolactin promoter, and at the same time longer durations of ON-periods for an artificial promoter, while Viñuelas et al. (2013) report mainly an increase in burst frequency upon HDAC inhibition. It seems that each gene uses different combinations of chromatin marks and promoter-specific transcription factors to define transcriptional states. This raises the question of how my results on the *GREB1* promoter would generalize to other genes. Even among E_2-dependent genes only a subset is repressed upon inhibition of HDAC activity, while others increase their expression (Reid et al. 2005). It seems that chromatin marks can achieve different outcomes at different genes, depending on promoter sequence and accessibility for transcription factors or co-factors.

In the model of estrogen-dependent promoter cycling, in which many biochemical steps lead to gene activation (Métivier et al. 2003), the inhibition of a single step would be detrimental for expression, largely independent of where in the cycle inhibition occurs. Given the multiple enzymatic activities involved at estrogen-dependent promoters, it is likely that inhibitors of histone methylation or DNA methylation would reveal further molecular characteristics that define transcriptional states. However, the nature of experiments in which enzymes are inhibited that act on multiple substrates precludes the identification of exact reactions. Likewise, the combinatorial complexity of chromatin marks is not accessibly through inhibitors. Additional, target-specific perturbations would be needed to define the molecular state of transcriptionally permissive chromatin. For example, one could mutate amino acid residues in target proteins such that they cannot be post-translationally modified. In addition, knockdown or knockout of genes, e.g. bromodomain-containing genes, would be more specific in identifying molecular targets of inhibitors. To link transcriptional output with specific chromatin modifications at the promoter it would be necessary to measure both aspects simultaneously in the same cell at a single locus. Chromatin modifications can be assessed at a specific locus using a combination of *in-situ* hybridization and proximity ligation (Gomez et al. 2013). In combination with nascent RNA imaging one could correlate promoter chromatin with transcriptional activity. However, this technique is limited to snapshot analyses, as it requires fixation of cells. Recently, a fluorescence complementation based method was developed, which provides an approach to detect chromatin modifications at specific sites in live cells (Lungu et al. 2017). Simultaneous live-imaging of promoter chromatin and associated RNA output is therefore within reach and will provide exciting insights into gene regulation.

3.3 Estrogen modulates frequency of transcriptional bursts

Transcriptional bursts achieve a desired expression level by turning two “dials”: either the number of bursts over time, i.e. the burst frequency, is varied, or the RNA production from each burst, i.e. the burst size, is altered. It was one aim of this thesis to reveal the regulatory principle in endogenous estrogen-mediated gene activation.

In case of *GREB1*, estrogen titration experiments revealed that expression is changed through modulating the frequency of gene activation. This is apparent through shortening of time intervals between bursts with increasing E_2 concentration, as shown by direct extraction from the raw fluorescence trajectories (Figure 19C). Furthermore, the posterior distributions of individual (Figure 30A) and global fits (Figure 32C), as well as response-times after E_2 starvation in single cells and cell populations (Figure 33), supported this finding. Frequency modulation was already postulated decades ago for nuclear receptor-mediated gene regulation (Ko et al. 1990) and recently confirmed by nascent transcript imaging of an insect hormone-controlled transgene (Larson et al. 2013). The results of my thesis add another example of direct observation of frequency-modulated transcription, but for a physiological signaling pathway at an endogenous locus within an unperturbed chromatin environment. Therefore, I provide strong evidence that frequency modulation is indeed a relevant regulatory mechanism. A common theme for nuclear receptor-mediated transcription seems that they promote the formation of a transcriptionally competent promoter state, thereby, controlling the kinetics of gene activation, but not inactivation. It seems plausible that a stimulus increases the probability that a promoter becomes permissive and once this commitment is made RNAs are produced. An altered frequency of activation leads to shorter response-times at high stimulus levels (Figure 33), but at the same time limits the amount of energy-dependent chromatin changes when stimulus is low. Interestingly, this regulatory principle further entails that expression noise decreases with increasing stimulus (Figure 36A). As I reveal the regulatory mechanism of estrogen, this thesis provides the basis for further studies on the gene-regulatory effect of pharmacological intervention targeting ERα.

Again, the question arises which specific molecular event characterizes the rate-limiting step in gene activation that is controlled by estrogen. The simplest biochemical explanation is that this transition coincides with the binding of the ligand-receptor complex to the response elements on the DNA. Higher concentrations of E_2 would shift the binding equilibrium such that ERα shows longer residence times or more frequent binding, the latter coinciding with burst frequency modulation. Residence times of nuclear receptors were measured by photobleaching on promoter arrays (summarized by Darzacq et al. 2009) and by single-molecule tracking (Groeneweg et al. 2014) and are reported to be less than 10 seconds. This is much shorter than the measured ON-times in *GREB1* transcription, which have a mean duration of 10 minutes when extracted from the raw data (Figure 19C) and are also below the mean ON-time from parameter posteriors of SMC ABC of about one minute (Figure 30A and 32C). It seems more likely that ERα binds frequently to DNA, but these interactions are short-lived and most often unproductive. This means that only

few of them initiate further biochemical reactions in the transition to a transcriptional active template (Métivier et al. 2006), similar to studies in yeast (Karpova et al. 2008, Poorey et al. 2013). Increased binding frequency or duration of ligand-receptor complexes would then increase the chance that such a productive event happens. However, the nature of the molecular event that characterizes a productive interaction remains unclear. It is tempting to speculate that chromatin plays a role because histone posttranslational modifications could persist even when the enzymes that set them only transiently interact in a "hit-and-run" mechanism (McNally 2000, Rigaud et al. 1991). Indeed, I found that histone acetylation is important for proper *GREB1* RNA production from an active state. Another possibility for persisting changes after short-lived interactions is chromatin remodeling, for example through removal of nucleosomes, with subsequent assembly of the transcription machinery. This step was found to be rate-limiting for the expression of the *PHO5* gene in yeast (Mao et al. 2010).

The role of transient transcription factor interactions with DNA in determining burst frequency are supported by studies that suggest that enhancers control burst frequency. Enhancers are short distal regulatory elements, which contain binding sites for diverse transcription factors and it has been proposed decades ago that they increase the probability of expression rather than its level (Walters et al. 1995, Weintraub 1988). Recent studies provide direct evidence for frequency regulation by live-cell imaging (Fukaya et al. 2016), and suggest that chromatin looping is important for mediating these effects (Bartman et al. 2016). *GREB1* has indeed a large enhancer region with multiple estrogen response elements, which loops to the promoter in an estrogen-dependent manner (Deschênes et al. 2007, Fullwood et al. 2009, Sun et al. 2007). As such, the enhancer regulation of *GREB1* bursting requires chromatin looping to the promoter, which could also be a rate-limiting step in gene activation.

A regulatory mechanism in which the burst frequency is regulated by the concentration of active ligand-receptor complexes means that the burst size can be used for gene-specific tuning of expression. In this scenario, two genes would respond with the same dose-dependency but could still achieve different mean expression levels. Indeed, burst frequency and burst size can be tuned independently: Senecal et al. (2014), for example, demonstrated that transcription factor concentration affects burst frequency, while the duration of DNA binding and strength of the activator domain affect burst size. In line with these results, Suter et al. (2011) described that burst size depends on number and affinity of transcription factor binding sites. In this thesis, I further identified the importance of protein (and likely histone) deacetylation and acetyl binding in the regulation of burst size. Hence, the locus-specific chromatin environment is also regulating RNA production from a permissive promoter state. It would be interesting to study other estrogen-dependent genes with distinct expression levels to reveal whether at a given E_2 concentration, target genes show similar burst frequencies but different burst sizes. Results from Larson et al. (2013) hint at such an effect. Different integration sites of a transgene showed five-fold difference in transcription, while the steroid hormone-controlled burst frequency remained unaltered. Furthermore, the analysis of different genes might reveal to what extent the en-

hancer configuration influences burst frequency. Taken together, transcriptional bursting provides a means for uncoupling gene-specific expression level and signal-dependency.

3.4 Chromatin control of intrinsic expression noise

The possibility to control burst size and burst frequency independently is also important when regulation of noise in gene expression is desired. Low levels of noise are for example required for a housekeeping gene, whose expression cannot drop below a critical level. In contrast, transcription factors that regulate differentiation might be noisier to provide stochastic cell fate decisions. Independent control of size and frequency of bursts allows to achieve the same expression level through different combinations these parameters. Because a given combination of burst frequency and burst size entails a specific intrinsic noise level, such dual control allows gene-specific noise regulation (Bar-Even et al. 2006). Indeed, burst size and burst frequency were observed to differ between genes (Dey et al. 2015, Molina et al. 2013, Suter et al. 2011) and to depend on the genomic location (Dar et al. 2012, Skupsky et al. 2010), suggesting that cells precisely control noise levels.

In case of my model system of signal-dependent transcriptional regulation, I observed an inverse noise-mean scaling in *GREB1* transcription (Figure 36), which results from changes in burst frequency across different levels of induction. Hence, noise is larger at low concentrations of estrogen and decreases at higher concentrations. It remains unclear whether frequency modulation passively entails this noise behavior or whether evolution preferred this specific noise-mean scaling and hence, frequency modulation was chosen as a regulatory principle. Clarifying such evolutionary questions is difficult. I could shift the specific noise-mean trajectory in estrogen-dependent transcription to lower noise levels through burst size reduction after HDAC inhibition. As such, I revealed a role for chromatin in specifying noise in gene expression. In fact, other studies also reported the dependence of gene expression noise on chromatin, for example, higher noise from repressed chromatin (Dey et al. 2015), altered burst sizes depending on the presence or absence of nucleosomes (Hornung et al. 2012), or changes upon treatment with inhibitors of chromatin modifying enzymes (Viñuelas et al. 2012). It is possible that the observed effect of HDAC inhibition is limited to a subset of estrogen-responsive genes, as effects on other genes are diverse. Often, HDAC inhibition leads to transcriptional upregulation, for example through increased burst sizes (Harper et al., 2011), but it can also be neutral, as intrinsic noise in Nanog expression is not affected (Ochiai et al., 2014).

Chromatin-mediated burst size control can alter gene expression independent of the burst frequency, but with a different noise-mean scaling: It is mainly the frequency of bursts that tunes noise levels, and burst size has only limited effects (Hornung et al. 2012). For estrogen-dependent transcription, in which the stimulus determines the burst frequency, the noise level would be determined by the stimulus such that noise and mean would not be independently controllable anymore. Further control of intrinsic noise can be achieved with additional states in the promoter cycle. Extra states produce more precise durations of active and inactive phases, which result in lower noise levels (Schwabe et al. 2012, Zhang et al. 2012). *GREB1* only has two rate-limiting steps in its promoter cycle and a two-state

model produces the highest intrinsic noise level. Hence, the *GREB1* promoter does not seem to be optimized for low expression noise.

In summary, promoter cycle topology with orthogonal control of burst size and frequency results in a specific noise level to occur for each gene at the level of nascent transcription. Post-transcriptional processes, such as RNA export, stability, and translation could then further modulate and reduce noise during propagation to protein levels (Stoeger et al. 2016, Thattai & van Oudenaarden 2001).

3.5 Extrinsic noise contribution to cell-to-cell variability

In the previous section, I discussed cellular control of intrinsic noise caused by transcriptional bursting itself. In addition, I observed strong cell-to-cell variability in long-term transcriptional output that is apparent when intrinsic noise is averaged out. Hence, burst parameters are additionally modulated by the cellular state (Figure 39). This effect had to be considered during stochastic modeling and indeed provided a better description of the data (Figure 26).

The first question that arises is how cellular state modifies burst characteristics to achieve differences in transcriptional output. As I directly observed the earliest phase of gene expression, posttranscriptional processes could be immediately ruled out as a potential source. Analysis of time-course features (Figure 20B) and model fitting (Figure 30B) revealed that the initiation rate and elongation kinetics differed between cells and caused the unequal transcriptional output. Indeed, initiation and elongation have been described to vary between cells. Annibale & Gratton (2015) for example, showed that elongation rate from a transgene array can vary ten-fold between cells, while das Neves et al. (2010) showed similar variation in elongation throughout the nucleus using photobleaching experiments of RNA PolII. It should be noted however, that the fitted *GREB1* elongation kinetics, which are captured in the rate k_{elong}, are not only determined by polymerase progression, but also by post-initiation processes such as polymerase pausing, splicing, or transcript retention, and that these additional steps contribute as well to the observed variability in elongation kinetics. In accordance with my observations, also the initiation rate was already reported to differ between cells and to account for most of the observed extrinsic noise in yeast (Sherman et al. 2015).

Initiation and elongation are regulated by the cellular state, but how this is achieved mechanistically remains unclear. A local, chromatin-dependent process (Jonkers & Lis 2015) can be excluded, because two alleles within the same cell correlate in long-term transcriptional output and in initiation rate (Figure 22B). This rather hints at a diffusible *trans*-acting factor that influences initiation and elongation rates globally. Cell cycle, volume, metabolic state, upstream signaling, and microenvironment are candidate mechanisms for global gene regulation (Battich et al. 2015, Padovan-Merhar et al. 2015, Skinner et al. 2016, Stewart-Ornstein et al. 2012). Battich et al. (2015) indicate that steady-state RNA levels are largely predictable for most genes, when image-based features such as nuclear and cellular morphology, mitochondrial content, or population context are consid-

ered. Hence, the cellular state strongly contributes to and determines population heterogeneity. Even though such analyses of RNA numbers do not reflect instantaneous gene activity because of buffering by posttranscriptional processes and the presence of multiple alleles, their study highlighted that morphological features correlate with transcript levels in a gene-specific manner.

One morphological feature that has been described to affect gene bursting is cellular volume. The burst size of multiple genes was observed to scale with cellular volume (Kempe et al. 2015, Padovan-Merhar et al. 2015). This suggests that differences in volume might contribute to the observation that total RNA output is correlated between *GREB1* alleles in the same cell. Furthermore, the cell cycle has a major impact on transcription (Padovan-Merhar et al. 2015, Skinner et al. 2016, Zopf et al. 2013). For instance, gene output has to be adjusted to varying copy number after replication. With respect to estrogen signaling, cell cycle dependence might further arise because ERα is regulated on RNA and protein levels throughout the cell cycle (JavanMoghadam et al. 2016, Vantaggiato et al. 2014). However, the impact of cell cycle in my experiments is probably minimal because only cells were analyzed which showed a transcription site, originating from a non-replicated allele. Hence, all analyzed cells are most likely in G0/G1 or early S phase of the cell cycle, but not in in late S or G2. Furthermore, in-silico synchronization in cell cycle to the time of cell division (Figure 23) revealed that cells still show heterogeneous RNA output when they are all in early G1 phase of the cell cycle. As cell cycle does not seem to contribute to extrinsic variably in *GREB1* expression, it would be interesting to record cellular volume, population context, and further features during live-cell imaging experiments and study their specific influence on *GREB1* initiation and elongation.

How cellular state mechanistically achieves changes in transcriptional kinetics remains unclear. Recent evidence links mitochondrial content to transcription elongation (das Neves et al. 2010, Johnston et al. 2012): Mitochondria supply at least some of the ATP that is needed for cellular anabolism, e.g. RNA production. Hence, their abundance can influence the rate of polymerase elongation. As cells divide, mitochondria could be unequally partitioned between daughter cells and could cause stochastic differences in energy content, and hence, transcriptional output. As such, ATP content would be a *trans*-acting factor, which can regulate gene output globally, providing one possible link of a morphological feature to transcription. Other *trans*-acting factors could be specific upstream signaling components (Filippi et al. 2016, Gandhi et al. 2011, Klein et al. 2015, Ochiai et al. 2014, Stewart-Ornstein et al. 2012), e.g. fluctuations in ERα levels, or the general transcription machinery, e.g. concentration of polymerases. In line with fluctuations in specific upstream components, Klein et al. (2015) and Stewart-Ornstein et al. (2012) observed that co-regulated genes show correlated mRNA expression in individual cells and Sigal et al. (2006) observed the same on protein levels. Similar fluctuations could hold for the level of ERα in my experiments, with cells showing high *GREB1* output having higher ERα levels than low expressers do. However, the question arises whether ERα levels would rather influence burst frequency and not burst size. Simultaneous labeling of two genes with different regulatory inputs in the same cell, or simultaneously meas-

uring ERα levels in the reporter cell lines, would further define and reveal the scope of extrinsic noise sources. Many more factors could contribute to extrinsic variability, as all proteins are subject to stochastic fluctuations and random partitioning during divisions (Huh & Paulsson 2011). The differences in transcriptional output and cellular state seem to be stable within the relatively long imaging timeframe of 750 minutes, in line with the timescale of changes in protein levels observed after cell division (Sigal et al. 2006). In combination with gene regulatory networks, such fluctuations could lead to meta-stable expression states, which manifest themselves as extrinsic variability.

I speculate that cells contain multiple "dials" for transcriptional control. Burst frequency, as determined by the binding of active receptor-ligand complexes to response elements, is used to adjust RNA production to external stimuli. The specific expression level and noise can then be fine-tuned at individual gene loci through burst size modulation by *cis*-regulatory elements on the DNA with recruitment of distinct co-regulators, and the local chromatin environment at the promoter, while global modulation through *trans*-acting factors allows simultaneous adjustments in many genes depending on cellular state, that is, volume or microenvironment. Timing of signal input (stable vs. oscillating) and localization of transcription factors can add additional layers of control (Lin et al. 2015, Sonnen & Aulehla 2014). As such, transcriptional regulation can be abstracted as multiple independent layers that allow integration of various regulatory inputs.

3.6 Buffering and consequences of expression noise

Transcriptional noise has been the subject of many scientific studies over the past years. It is now appreciated that cells precisely control the amount of uncertainty in protein abundance through buffering or amplification of intrinsic noise and that variation is exploited in cellular decision-making.

Directly observing nascent RNAs enabled me to observe stochastic effects in transcription itself, unperturbed by post-transcriptional regulatory processes. The amount of intrinsic, bursting-related noise dominated over extrinsic noise across experimental conditions (CV^2_{int} = 0.5–3, CV^2_{ext} = 0.1–0.6, Figure 36), as short-term fluctuations in transcriptional activity were pronounced. As discussed before, *GREB1* does not seem to be optimized for low-noise expression, as its promoter cycle contains only two states and transcription is regulated through frequency modulation. The question arises whether *GREB1*-dependent cellular responses to estrogen are at all affected by fluctuating nascent transcription or to what extent such noise is buffered. Indeed, various processes must occur between the synthesis of nascent RNA and mRNA translation and these may act to buffer protein production. Nuclear export (Battich et al. 2015, Halpern et al. 2015) and compartmentalization in general (Stoeger et al. 2016) buffers bursts in RNA production when export is slow. Furthermore, long mRNA and protein half-lives effectively reduce fast temporal fluctuations in RNA levels (Paulsson 2005). The mRNA half-live of *GREB1* is 4.4 hours in mouse embryonic stem cells (Sharova et al. 2009), which is much longer than OFF-times at intermediate estrogen stimulus, hence, efficient buffering of noise occurs on the RNA level.

While bursting from a single allele causes strong variability, the presence of uncorrelated bursting from multiple alleles (Figure 21, Gandhi et al. 2011, Levesque & Raj 2013, Skinner et al. 2016) reduces noise (Raser & O'Shea 2004). Similarly, homologs or gene duplications as well as redundancies in cellular pathways that rely on the expression of distinct genes, increase the apparent gene number, and robustly buffer fluctuations. The multitude of buffering mechanisms only allows for minimal propagation of intrinsic noise to the protein level. Nevertheless, differences in transcriptional output arising from extrinsic variability (up to ten-fold difference in RNA production over 12 hours, Figure 18A) are temporally stable, suggesting that they propagate through to protein levels, even when short-time bursting is efficiently buffered by cellular systems.

3.7 Quantitative insight into gene regulation

The normalization of measured fluorescence intensities to single RNA molecules in live-cell imaging and smRNA FISH experiments (Figure 16) allowed me to express transcriptional activity in terms of absolute RNA numbers. Model fitting then revealed that on average eight RNAs are produced during a transcriptionally active interval, largely independent of the E_2 concentration (Figure 32C). This number is in quantitative agreement with other studies that report a burst size of 5–30 RNAs/burst for endogenous genes (Suter et al. 2011) with up to 100 RNAs/burst for strong promoters of housekeeping genes (Dar et al. 2012). The maximum initiation rate I observed was 10 RNAs/min (Figure 19B). At an elongation rate of 3.5 kb/min this would result in a minimal polymerase spacing of about 350 bp—about the size of two nucleosomes. This value seems at least realistic for separating polymerases. Tantale et al. (2016) suggest that it would be energetically favorable if polymerases travel together in "convoys" to avoid excessive removal of supercoiling while reporting a similar value of 280 bp for polymerase spacing within such a convoy.

The high quality of my single-cell fluorescence measurements, with manual inspection and correction of each trajectory, results in precise temporally resolved data of transcriptional pulses. The determined OFF-times for *GREB1* varied between 10 and 500 minutes throughout induction levels (Figure 32C), in agreement with reports for other mammalian genes (Dar et al. 2012, Larson et al. 2013, Molina et al. 2013). These studies report ON-times that are usually within the range of 10–60 minutes. This is similar to the ON-times that were directly extracted from the raw trajectories (mean of ~10 minutes, Figure 19C) but remarkably longer than in the posterior distribution after SMC ABC (mean of 0.56 min, Figure 32C). Why the SMC ABC algorithm preferred such short ON-times is unclear as it is not apparent how this ON-time is reflected in the features that were used to calculate the distance metric.

3.8 Consequences of noise on heterogeneity in cancer

Non-genetic expression heterogeneity has widespread impact on tumor progression and drug resistance (Brock et al. 2009, Capp 2005, Cohen et al. 2008, Kreso et al. 2013, Sharma et al. 2010). Stable adaptations of the cellular expression program, e.g. in oncogenes and tumor suppressors, can for example result from chromatin-mediated, epigenet-

ic alterations, without the need for genomic mutations. Also, gene expression is stochastic and induces fluctuations in protein levels, such that manifestation through gene regulatory networks may create subpopulations of cells that are able to escape therapy (Paek et al. 2016, Roux et al. 2015). Understanding determinants of cellular noise is therefore critical. This thesis revealed how expression noise from the endogenous, estrogen-regulated *GREB1* locus arises and can be controlled at the level of nascent transcripts. GREB1 is an important growth regulator in estrogen-dependent breast cancer (Rae et al. 2005). Hence, variability in protein levels might contribute to heterogeneous growth in cancerous tissue and have an impact on therapeutic intervention. I previously discussed that cells contain noise buffering systems, but alterations in these systems with effects on critical parameters, e.g. protein and mRNA decay rates, could increase general noise levels. Furthermore, as I found that *GREB1* noise can be tuned through chromatin-dependent changes in burst size, chromatin-modifying enzymes are further targets that could be exploited by cancer cells for noise manipulation. Increased noise levels would then increase the probability of extreme expression levels leading to more diverse population behavior.

ERα regulates gene expression through frequency-modulation with profound implications on noise-mean scaling (Figure 36). ERα antagonists, such as fulvestrant (ICI 182,780) or tamoxifen, which are used in therapy of breast cancer, downregulate gene expression by modulating levels of ERα or active ligand-receptor complexes, respectively. This suggests that they achieve downregulation along the same noise-mean trajectory as E_2, leading to higher noise levels when gene expression is inhibited. Hence, while global gene expression decreases, variability is high and individual cells could still randomly express *GREB1*. In addition, the inhibition by ERα antagonists is incomplete, as 20–40 % of cells still showed *GREB1* transcription sites, even at high inhibitor levels (Figure 35). In comparison, inhibition of acetylation-related processes was less variegated, as only 10 % of cells still transcribed. The fact that HDAC inhibition downregulated *GREB1* expression in more cells and at the same time also reduced noise, suggests that therapeutic success might be higher when ERα antagonists and HDAC inhibitors would be applied in combination. Follow-up experiments with co-treatments are needed to confirm this intriguing prediction.

3.9 Future perspectives

This thesis revealed that extrinsic noise is the dominant source of cell-to-cell variability at long time-scales, even when bursts produce strong intrinsic noise. As such, it is the cellular state that determines long-term transcriptional output, for example in growth regulators, and contribution to tissue heterogeneity is apparent. It will become important to understand heterogeneous cellular states within healthy and cancerous tissue and their contribution to gene output and cellular behavior.

Unraveling how cellular state impacts on gene transcription to contribute to expression heterogeneity will require further experiments. As previously suggested, one could measure additional features of single cells in combination with transcriptional output. Imaging-based analysis could be used to determine aspects of morphology, cell cycle, and micro-

environment. Recent developments in single-cell analysis enable live-cell imaging followed by single-cell RNA sequencing (Lane et al. 2017). Thereby, the whole transcriptome is accessible for the unbiased analysis of cellular state with its contribution on long-term transcriptional responses. In addition, single-cell approaches in proteomics (Budnik et al. 2017, Lombard-Banek et al. 2016, Newman et al. 2006) could provide analysis on the contribution of proteins and their posttranslational modifications to cellular state. Even though chromatin did not seem to be directly involved in mediating effects of cellular state, I am convinced that single-cell approaches to epigenetics will further improve our understanding of expression and noise regulation. Chromatin binding by ChIP (Rotem et al. 2015), chromatin structure by ATAC-Seq (Buenrostro et al. 2015) and Hi-C (Nagano et al. 2013), and DNA methylation by bisulfite-Seq (Smallwood et al. 2014) are accessible at the single cell level already today. Combination with expression analysis will provide exciting insights into determinants of gene activity on the single-cell, single-promoter level. Because extrinsic factors likely affect multiple genes, it will be necessary to analyze multiple genes, ideally in the same cell. Advances in labeling techniques for dynamic observation of nascent RNAs (Abudayyeh et al. 2017, Nelles et al. 2016) and genomic loci (Ochiai et al. 2015) without the need for genomic alterations circumvent the time-consuming step of establishing cell lines and will aid the analysis of multiple genes to differentiate upstream effectors of expression noise.

Targeting gene expression noise and cell-to-cell variability will become increasingly valuable for cancer therapies and further research is required to gain a mechanistic understanding of origins of variability and how treatments affect noise.

4 Materials and Methods

4.1 Materials

4.1.1 Chemicals

Table 1: Chemicals used in this study.

Name	Supplier	Catalog number
17β-estradiol	Sigma-Aldrich	E8875
Actinomycin D	Sigma-Aldrich	A1410
C646	Sigma-Aldrich	SML0002
Catalase	Sigma-Aldrich	C3155
DAPI	Sigma-Aldrich	D9542
DMSO	Sigma-Aldrich	D8418
DRAQ5	eBioscience	65-0880-92
Flavopiridol	Sigma-Aldrich	F3055
Formamide, deionized	Applichem	A2156
Glucose	Sigma-Aldrich	D9434
Glucose oxidase	Sigma-Aldrich	G0543
4-Hydroxy-Tamoxifen	Sigma-Aldrich?	579002
ICI 182,780	Selleckchem	S1191
MG132	Sigma-Aldrich	M7449
Paraformaldehyde	Sigma-Aldrich	16005
PFI-1	Sigma-Aldrich	SML0352
Puromycin Dihydrochloride	Thermo Fisher Scientific	A1113803
Sodium butyrate	Sigma-Aldrich	303410
Trichostatin A	Sigma-Aldrich	T1952
Trolox	Sigma-Aldrich	238813

4.1.2 Buffers and solutions

Table 2: Buffers and solutions used in this study.

Name	Supplier	Catalog number
Charcoal-stripped FBS	Sigma-Aldrich	F6765
DNAzol	Thermo Fisher Scientific	10503-027
DMEM	Lonza	BE12-614F
DMEM without phenol red	Thermo Fisher Scientific	31053-028
FBS-Gold	GE Healthcare	A15-151
L-Glutamine	Lonza	BE17-605E
PBS	Lonza	BE17-612F
Penicillin/Streptomycin	Lonza	BE17-602E
Stellaris ® RNA FISH wash buffer A	LGC Biosearch Technologies	SMF-WA1-60
Stellaris ® RNA FISH wash buffer B	LGC Biosearch Technologies	SMF-WB1-20
Stellaris ® RNA FISH hybridization buffer	LGC Biosearch Technologies	SMF-HB1-10
TRIzol	Thermo Fisher Scientific	15596018
Trypsin-EDTA (0.25 %)	Thermo Fisher Scientific	25200-056

4.1.3 Enzymes and Markers

Restriction endonucleases were purchased from New England Biolabs. As marker for DNA size, a GeneRuler™ 1 kb ladder from Thermo Fisher Scientific was used.

4.1.4 Kits

Table 3: Kits used in this study.

Name	Supplier	Catalog number
FuGENE® HD Transfection Reagent	Promega	E2311
In-Fusion® HD Cloning Plus	Clontech	638909
Lipofectamine® LTX	Thermo Fisher Scientific	A12621
Plasmid Plus Midi Kit	Qiagen	12945
Q5® High-Fidelity DNA Polymerase	New England Biolabs	M0491
QIAquick Gel Extraction Kit	Qiagen	28706
QIAquick PCR Purification Kit	Qiagen	28106
QIAprep Spin Miniprep Kit	Qiagen	27106
SuperScript® II RT	Thermo Fisher Scientific	18064014
SYBR Green PCR Master Mix	Thermo Fisher Scientific	4364344
TrueStart Hot Start *Taq*	Thermo Fisher Scientific	EP0613

4.1.5 Laboratory equipment

Table 4: Machines used in this study.

Name	Description	Company
ARIA III SORP	Cell sorter	Becton Dickinson
DeltaVision™ Elite	Live-cell widefield microscope	GE Healthcare
FORMA Steri-Cult™	Cell culture incubator	Thermo Fisher Scientific
MDF-C2156VAN-PE	-150 °C freezer	Sanyo
NanoDrop™ 2000	UV-Vis Spectrophotometer	Thermo Fisher Scientific
Opera Phenix	High-content screening microscope	Perkin Elmer
SterilGARD III	Biological Safety Cabinet	The Baker Company
Imcubator cool	Temperature control	Imsol
CO_2 controller	CO_2 mixer	Leica
TC10	Cell counter	BioRad
TProfessional TRIO	PCR machine	Analytik Jena
ViiA™ 7	Real-time PCR machine	Thermo Fisher Scientific

Table 5: Essential plastic ware used in this study.

Name	Supplier	Catalog number
96-well SensoPlate™ Plus	greiner	655891
LightCycler® 480 Multiwell Plate 384	Roche	04729749001
PYREX® Cloning Cylinder, 6 x 8 mm	VWR	22877-252
µ-Slide VI 0.4 ibiTreat	ibidi	80606

4.1.6 Oligonucleotides

All Oligonucleotides were ordered from Sigma-Aldrich with desalting purification. Oligonucleotides for cloning were reverse phase cartridge purified.

Table 6: Oligonucleotides used for cloning in this study.

Name	Sequence
inF_PCPlinker-GPF-for	GCTGTACAAGTCCGGACTCAGATCTCGAGCTCAAGCTTCGAATTCTGTGAG-CAAGGGCGAGGAGCT
inF_GFP-rev	TTATCTAGATCCGGTGGATCCCGGGTTACTTGTACAGCTCGTCCATGCCG
inF_K2L-PCP-for	ACCGTCAGATCCGCTAGTGCCCGCCATGGGCCCAAA
inF_GFP-PCP-rev	CTGAGTCCGGACTTGTACAGCTCGTCCATGCCG
tdPCP-NotI-for	ATTTAGCGGCCGCCATGTCCAAAACCATCGTTCTTTCGGTC
tdGFPSV40-ClaI-rev	GCCTCATCGATGTTAAGATACATTGATGAGTTTGGACAAACC
pSB_bb_ClaI_for	AGGCAATCGATCCCAAGTTAAACAATTTAAAGGCAATGC
pSB_bb_SpeI_rev	AACCTACTAGTCCAAGCTGTTTAAAGGCACAGTCAA
gRNA_GREB1_in2_t	CACCGTCTCACACAAGCTCAGTGTG
gRNA_GREB1_in2_b	AAACCACACTGAGCTTGTGTGAGAC
gRNA_GREB1_ex2_t	CACCGTTTCACCTTCTACCTTGCG
gRNA_GREB1_ex2_b	AAACCGCAAGGTAGAAGGTGAAAC
gRNA_GREB1_ex33_t	CAACGCATGGAAGCACACGAATGGC
gRNA_GREB1_ex33_b	AAACGCCATTCGTGTGCTTCCATGC
inF_GREB1_in2_L_PP7_rev	ATTCGTTTAAACCTGCAGGAGATCTCACACAAGCTCAGTGTGGAGCCCCTT-GAAGACAA
inF_GREB1_in2_L_pUC_for	AGTCGACCTGCAGGCATGCATCAGTAGGAGAAAGGAAGAGGAC
inF_GREB1_in2_R_PP7_for	CTCGCGAAGGGCGAATTCGCTGATAGTTGCTTATTCTAGCAGGATGCATTTCTGT
inF_GREB1_in2_R_pUC_rev	CTATGACCATGATTACGCCAAGGTGGGCAGAAAGCAACTA
inF_loxP_GREB1_R_for	TTATCTGATAGTTGCTTATTCTAGCAGGATGC
inF_loxP_NotSpe-GREB1_L_rev	ATAACTTCGTATAATGTATGCTATACGAAGTTATAAGCGGCCGCTAC-CAACTAGTGATCTCACACAAGCTCAGTGTGG
inF_BFP_IRESPuro_for	AAGTCCAAGCTGTAATGCATCTAGGGCGGCCAA
inF_loxP_invPuro_rev	ATTATACGAAGTTATCCATAGAGCCCACCGCATC
inF_loxP_invCMVBFP_for	GCAACTATCAGATAACTTCGTATAATGTATGCTATACGAAGTTATAACCG-TATTACCGCCATG
inF_IRES_BFP_rev	TTACAGCTTGGACTTGTACAGCTCGTC
inF_GREB1_ex2_L_pUC_for	CGACGGCCAGTGAATTCAGTGGTTCTCTGTGATTTTTGGG
inF_GREB1_ex2_L_loxP_rev	TACCAACTAGTGATCAAGGTGAAACAGCTGCAAGGA
inF_GREB1_ex2_R_loxP_for	TCTGATAGTTGCTTATGCACCGAATCTGAGATGCCA
inF_GREB1_ex2_R_pUC_rev	TCGACTCTAGAGGATCACTGAGGGGTTACTTCTGA
inF_GREB1_ex33_L_pUC_for	CGACGGCCAGTGAATTAAGGTGGGGTAGGTAGGGTG
inF_GREB1_ex33_L_loxP_rev	TACCAACTAGTGATCGGAAGTAAATCTTTGTTCCTCGTG
inF_GREB1_ex33_R_loxP_for	TCTGATAGTTGCTTATGTGCTTCCATGGACAAACCTG
inF_GREB1_ex33_R_pUC_rev	TCGACTCTAGAGGATCAGTCTGTGAAGGGGTGAGCC
inF_BFP-Puro_SpeNot_for	GATCACTAGTTGGTAGCGGCC
inF_BFP-Puro_rev	TAAGCAACTATCAGATAACTTCGTATAATGTATG
inF-GAPDH-pUC-for	AAAACGACGGCCAGTGAATTTGCCTCCTGCACCACCAAC
inF-GAPDH-GREBwt-rev	AAGGTGAAACCCACAGTCTTCTGGGTGGCAG
inF-GREBwt-GAPDH-for	AAGACTGTGGGTTTCACCTTCTACCTTGCGTGGA
inF-GREBwt-ki-rev	AGAGACGAAGGCAAGACCTCTTCAAAGCGTGTC
inF-GREBki-wt-for	GAGGTCTTGCCTTCGTCTCTGCTGAGCGAAGG
inF-GREBki-pUC-rev	CAGGTCGACTCTAGAGGATCTTAGGTACCTTAGGATCCCGCGAAG

Table 7: Oligonucleotides used for genotyping PCRs in this study. All primer pairs have been checked for specificity against the human genome using Primer-BLAST[1].

Name	Sequence	Location
GREB1_in1_for	CCATCCCTTCCCATCTGCAAG	Outside of homology arms of exon 2 knock-in
GREB1_in3_rev	TGCGGGTAACTTCAAGTCAAAG	
HR_PP7_rev	TACCTTAGGATCCCGCGAAG	5' of PP7 sequences
HR-CMV rev2	ACCGTAAGTTATGTAACGCGGAACT	In CMV promoter
PP7loxP_for2	TAGATCTCGCGAAGGGCGAA	3' of PP7 sequences
GREB1_ex33_for3	TGAGCACCAGCAGCATCTAA	Outside of homology arms of exon 33 knock-in
GREB1_ex33_rev2	AATGGGTGGCCTCTCAACTG	
GREB1_in2_for	CTAGAAGGTGGGAGACGCAC	Outside of homology arms of intron 2 knock-in
GREB1_in2_rev	CATACCAACGTGGAGCTGGA	
GREB1_in2_rev2	CCCCCTTGTTTTCCTGCACTAGA	

Table 8: Oligonucleotide pairs used for RT-qPCRs in this study. All primer pairs have been checked for specificity against the human transcriptome (Refseq mRNA) using Primer-BLAST[1].

Name	Sequence	Target	Amplicon size
GREB1_exin2_1_f	GTCCAACAACCTGGTGCC	*GREB1*, exon2-intron2	104 bp
GREB1_exin2_1_r	CAGATAAAAGCAACGTGCGTC		
GREB1_ex2_1_f	ACCTTCTACCTTGCGTGGAG	*GREB1*, wt (allele-specific)	129 bp
GREB1_ex2_1_r	CCTCTTCAAAGCGTGTCGTC		
GREB1_f	CCCATCTTTTCCCAGCTGTA	*GREB1*	117 bp
GREB1_r	ATTTGTTTCCAGCCCTCCTT		
GREB1_exin32_1_f	GCCGCTTTCCTCTGGATAAAC	*GREB1* intron32-exon33	82 bp
GREB1_exin32_1_r	GATGACACACAACGTCGCA		
HR_PP7_rev	TACCTTAGGATCCCGCGAAG	*GREB1*, PP7 (allele-specific)	107 bp
GREB1_ex2_4_f	TCTCTGCTGAGCGAAGGC		
GAPDH_f	CTGCACCACCAACTGCTTAG	*GAPDH*	108 bp
GAPDH_r	GTCTTCTGGGTGGCAGTGAT		

[1] https://www.ncbi.nlm.nih.gov/tools/primer-blast (2013-2016)

Table 9: Probes for single-molecule RNA FISH of intronic regions of *GREB1*. Probes were ordered from LGC Biosearch Technologies labeled with Quasar® 670.

Number	Sequence	Target
1	AGAAGTCTGCGGGTAACTTC	*GREB1* intron 2
2	CACAGTCTAGTTTCTCTCAG	*GREB1* intron 2
3	AAGCTGAGTATGCAACTGCT	*GREB1* intron 2
4	TTTTTATCTCAGAGCTTGGC	*GREB1* intron 2
5	TGTCCAACAGCAGATAAGGG	*GREB1* intron 2
6	AGTGCTTGGTTTAGGGTAAC	*GREB1* intron 2
7	CCAGGTTTGACTATGCAGAA	*GREB1* intron 2
8	CCCGACAAACAGACACTACG	*GREB1* intron 2
9	GATCAAGGGGTGTTCAGTAT	*GREB1* intron 2
10	AACAAGCTCGAGGGGACATT	*GREB1* intron 2
11	CAGACTCTCAGAAGGCATGA	*GREB1* intron 2
12	AATAAGGGAGACACATCCCA	*GREB1* intron 2
13	CACAGATGCCAACTATCAGT	*GREB1* intron 2
14	CCATGTCGGACACAGAAGAC	*GREB1* intron 2
15	CGTGGTTTTAGCAGAAGGTG	*GREB1* intron 2
16	GGCTAAGCAACCTAGGTTAA	*GREB1* intron 2
17	AGGATCTCAGGCTGTTGAAA	*GREB1* intron 2
18	GGTACACAGAATCGTACCTT	*GREB1* intron 2
19	ATTATGACTGTGTGTTGGGC	*GREB1* intron 2
20	ATGAAATTCTCTCAGCTCCG	*GREB1* intron 2
21	AACAGGGTCAGGTATGGTAT	*GREB1* intron 2
22	GTAGATGGTACCATAGTGTC	*GREB1* intron 2
23	TAACAAGCACTCAGCACGTC	*GREB1* intron 2
24	TAATCCAGAATGGGTTCCAC	*GREB1* intron 2
25	TTTAGGTTTGTCTCAGGAGG	*GREB1* intron 9
26	CTGATCAGGGGCTGAGCTAG	*GREB1* intron 9
27	GACAAGCTCCATCATTCTTG	*GREB1* intron 9
28	CATGGCAGGGAATCATTTCA	*GREB1* intron 9
29	CGTGGGTGATAACAGGAGAC	*GREB1* intron 9
30	TTGATGCTTCTACAGTGGTG	*GREB1* intron 9
31	TCAGCTGAGCCAAAAGTCTA	*GREB1* intron 9
32	AGGGTTTCCCATTGCAACAT	*GREB1* intron 9
33	ATTCTGGTGCCTACATATCA	*GREB1* intron 9
34	AACGTATTTAAGAGGGGCCT	*GREB1* intron 9
35	TTAACAGTAGGGTGCTTCTC	*GREB1* intron 9
36	GCGTCATGCTAAGGTCGAAA	*GREB1* intron 9
37	TGGAACTGCCTCGGTCATTG	*GREB1* intron 9
38	TTTCTTTTCCGAGGTTCCTG	*GREB1* intron 9
39	TGTGAGACGTAGACTTGCTG	*GREB1* intron 9
40	AGCAGAGCACGCCTGAGAAC	*GREB1* intron 9
41	GAGTCTTAAGGCCTCAGGAG	*GREB1* intron 9
42	CACAGTGACTTCATGACGTC	*GREB1* intron 10
43	TCAACCTCAACCTACTTTCA	*GREB1* intron 10
44	TTCATGACCTAAACTGACCC	*GREB1* intron 10
45	GGGACTATGAGAAAGAGCGA	*GREB1* intron 10
46	CGTGCTTACTGATGGACAGG	*GREB1* intron 10
47	AATCGGAGTCCAAGTTCTCA	*GREB1* intron 10
48	ACTATTCCCTCAATATAGGT	*GREB1* intron 10

Probes against exonic regions were obtained from LGC Biosearch Technologies labeled with Quasar® 570 (Catalog #VSMF-2158-5).

4.1.7 Plasmids

Table 10: Plasmids used in this study.

Name	Purpose	Origin
pHAGE-UbC-NLS-HA-tdPCP-GFP	Mammalian expression of tandem dimer PP7 coat-protein-GFP fusion	gift from Robert Singer (Addgene #40650)
pSB-ET-iE	Cloning vector with inverted repeats for Sleeping Beauty transposition; reverse tetracycline-controlled transactivator IRES-Puro	gift from Manfred Gessler
pCMV(CAT)T7-SB100X	Mammalian expression of hyperactive Sleeping Beauty transposase	gift from Zsuzsanna Izsvak (Addgene #34879)
pSB-Ubc-tdPCP-tdGFP	Mammalian expression of tandem dimer PP7 coat-protein fused to tandem dimer GFP with inverted repeats for Sleeping-Beauty transposition	this study
pX330-U6-Chimeric_BB-CBh-hSpCas9	Mammalian expression vector for *S. pyogenes* Cas9 and cloning site for chimeric guide RNA	gift from Feng Zhang (Addgene #42230)
pX330-GREB1-in2L	Cas9 and guide RNA for intron 2 of *GREB1*	this study
pX330-GREB1-ex2T	Cas9 and guide RNA for exon 2 of *GREB1*	this study
pX330-GREB1-ex33B	Cas9 and guide RNA for exon 33 of *GREB1*	this study
pCAG-Cre-IRES2-GFP	Mammalian expression of Cre recombinase	gift from Anjen Chenn (Addgene #26646)
pEBFP2-Nuc	Cloning template for eBFP2	gift from Robert Campbell (Addgene #14893)
pGLUE	Cloning template for IRES-Puro	gift from Randall Moon (Addgene #15100)
pCR4-24xPP7SL	Cloning template for PP7 stem-loop cassette	gift from Robert Singer (Addgene #31864)
pHR-GREB1-in2-24xPP7-LPIBCL	Template for homologous recombination; insertion of PP7 and selection cassette into intron 2 of *GREB1*	this study
pHR-GREB1-ex2-24xPP7-LPIBCL	Template for homologous recombination; insertion of PP7 and selection cassette into exon 2 of *GREB1*	this study
pHR-GREB1-ex33-24xPP7-LPIBCL	Template for homologous recombination; insertion of PP7 and selection cassette into exon 33 of *GREB1*	this study
p-mKate2-C	Cloning template	Evrogen, FP181
pUC19	General cloning vector	NEB, N3041S
pUC-qRT-GAPDH-GREB1ex2-wt-PP7	Reference for absolute quantification of gene expression in qPCRs	this study

Plasmids were cloned as outlined below. Parts that were changed during a cloning step were sequenced to ensure correctness.

pSB-Ubc-tdPCP-tdGFP

The eGFP ORF was amplified from pHAGE-UbC-NLS-HA-tdPCP-GFP using primers inF_PCPlinker-GPF-for and inF_GFP-rev. In a second PCR on the same template the tdPCP-GFP ORF was amplified with primers inF_K2L-PCP-for and inF_GFP-PCP-rev. The pmKate2-C vector was linearized with *Xma*I and *Kpn*I and used together with both PCR products in an In-Fusion reaction to yield the intermediate pmKate2-C-tdPCP-tdGFP. This vector was used as a PCR template to amplify the tdPCP-tdGFP ORF together with an SV40-polyadenylation sequence using primers tdPCP-NotI-for and tdGFPSV40-ClaI-rev. The PCR product was digested with *Age*I and *Cla*I and replaced the *Age*I-*Cla*I fragment from pHAGE-UbC-NLS-HA-tdPCP-GFP yielding pHAGE-UbC-NLS-HA-tdPCP-tdGFP. The complete expression cassette was cut with *Spe*I and *Cla*I and ligated with the

*Cla*I-*Spe*I digested PCR product that resulted from a PCR on pSB-ET-iE with the primers pSB_bb_ClaI_for and pSB_bb_SpeI_rev to produce pSB-Ubc-tdPCP-tdGFP.

pX330 with custom guide RNAs

Guide RNA sequences were ordered as complementary oligonucleotides (see Table 6), annealed, and inserted in between the *Bbs*I sites of pX330-U6-Chimeric_BB-CBh-hSpCas9.

pHR-GREB1 plasmids

Homology arms up- and downstream of the Cas9 cleavage site in intron 2 of *GREB1* were amplified from genomic DNA of MCF-7 cells using primers inF_GREB1_in2_L_PP7_rev and inF_GREB1_in2_L_pUC_for, as well as inF_GREB1_in2_R_PP7_for and inF_GREB1_in2_R_pUC_rev. pUC19 was linearized with *Hin*dIII and the 24xPP7 cassette was cut from pCR4-24xPP7SL using *Spe*I and *Not*I. All four fragments were assembled using In-Fusion cloning to yield pHR-GREB1-in2-24xPP7. This vector was used to amplify the homology arms with the vector backbone by PCR with primers inF_loxP_GREB1_R_for and inF_loxP_NotSpe-GREB1_L_rev. The selection cassette consists of a CMV promoter that drives expression of a bicistronic mRNA with an eBFP2 carrying a peroxisomal targeting sequence and a Puromycin resistance gene after an IRES sequence flanked by two loxP sites. It is derived by two PCRs. First, IRES-Puro and the bGH polyadenylation signal were amplified from pGLUE with primers inF_BFP_IRESPuro_for and inF_loxP_invPuro_rev. Second, the CMV-eBFP2 cassette was amplified from pEBFP2-Nuc with primers inF_loxP_invCMVBFP_for and inF_IRES_BFP_rev. All three PCR fragments were assembled using In-Fusion cloning to yield pHR-GREB1-in2-LPIBCL. The 24xPP7 cassette was cut from pCR4-24xPP7SL using *Spe*I and *Not*I and inserted into the *Spe*I and *Not*I sites to yield the final pHR-GREB1-in2-24xPP7-LPIBCL.

To exchange the homology arms of intron 2 of *GREB1* against sequences in exon 2 and exon 33, the respective homology arms were amplified from genomic DNA of MCF-7 cells using primers inF_GREB1_ex2_L_pUC_for and inF_GREB1_ex2_L_loxP_rev for the 5' arm of exon 2, inF_GREB1_ex2_R_loxP_for and inF_GREB1_ex2_R_pUC_rev for the 3' arm of exon 2, inF_GREB1_ex33_L_pUC_for and inF_GREB1_ex33_L_loxP_rev for the 5' arm of exon 33, inF_GREB1_ex33_R_loxP_for and inF_GREB1_ex33_R_pUC_rev for the 3' arm of exon 33. The two PCR products with the 3' and 5' homology arm for each construct were used together with a PCR product with primers inF_BFP-Puro_SpeNot_for and inF_BFP-Puro_rev on pHR-GREB1-in2-24xPP7-LPIBCL and an *Eco*RI + *Bam*HI linearized pUC19 plasmid in an In-Fusion reaction. The 24xPP7 cassette was cut from pCR4-24xPP7SL using *Spe*I and *Not*I and inserted into the *Spe*I and *Not*I sites of the assembled vectors to yield the final pHR-GREB1-ex2-24xPP7-LPIBCL and pHR-GREB1-ex33-24xPP7-LPIBCL.

pUC-qRT-GAPDH-GREB1ex2-wt-PP7

qPCR products were amplified using primers inF-GAPDH-pUC-for with inF-GAPDH-GREBwt-rev, inF-GREBwt-GAPDH-for with inF-GREBwt-ki-rev, and inF-GREBki-wt-for with inF-GREBki-pUC-rev from cDNA of E_2 treated MCF7–PCP_GREB1_ex2_c16 cells.

The three PCR fragments were inserted into an *EcoR*I + *BamH*I linearized pUC19 backbone using an In-Fusion reaction to yield pUC-qRT-GAPDH-GREB1ex2-wt-PP7.

4.1.8 Cell lines

Table 11: Human cell lines used in this study.

Name	Description	Origin
MCF7	human breast cancer cell line; ERα positive; parental cell line for all stable cell lines generated in this study	Edison T. Lui
MCF7–SBtdPCP-tdGFP	clonal cell line; stable Sleeping Beauty-transposase mediated integration of tdPCP-tdGFP cassette	this study
MCF7–noPCP_GREB1_in2_c1	clonal cell line; knock-in of 24xPP7-LPIBCL into intron 2 of *GREB1*; two alleles labeled	this study
MCF7–PCP_GREB1_in2_c1	clonal cell line; as above, with stable tdPCP-tdGFP expression	this study
MCF7–PCP_GREB1_ex2_c16	clonal cell line; stable tdPCP-tdGFP expression; knock-in of 24xPP7-LPIBCL into exon 2 of *GREB1*; one allele labeled	this study
MCF7–PCP_GREB1_ex2_c16_Cre	clonal cell line; stable tdPCP-tdGFP expression; knock-in of 24xPP7 in exon 2 of *GREB1*; selection cassette excised by Cre recombinase	this study
MCF7–PCP_GREB1_ex33_c3	clonal cell line; stable tdPCP-tdGFP expression; knock-in of 24xPP7-LPIBCL into exon 33 of *GREB1*; one allele labeled	this study

4.1.9 Software

Table 12: Software used in this study.

Name	Supplier	Version
Columbus	PerkinElmer	2.7.1
CRISPR design tool	Feng Zhang, http://crispr.mit.edu	October 2014
Harmony	PerkinElmer	4.1
ImageJ	Wayne Rasband	1.51
MATLAB	MathWorks	2015b
Office	Microsoft	2010
Primer-BLAST	Ye et al. 2012	-
Python	Python Software Foundation	2.7
R	R Core Team	3.3.0
softWoRx	GE Healthcare	6.5.2
Stellaris® RNA FISH Probe Designer	Biosearch Technologies	4.2
uTrack	Jaqaman et al. 2008	2.0

4.2 Cell culture

4.2.1 Maintenance, passaging and long-term storage of cells

All cell lines were maintained in Dulbecco's Modified Eagle Medium (DMEM) with 4.5 g/L Glucose (Lonza) supplemented with 10 % fetal bovine serum (FBS) (FBS-Gold, GE Healthcare), 1 % L-Glutamine (Lonza) and 1 % Penicillin/Streptomycin (Lonza) at 37 °C in a humidified atmosphere containing 5 % CO_2. The medium was replaced with fresh medium every 2–3 days. Cells were subcultured at a ratio of 1:3 to 1:6 when a confluency of 90 % was reached. For this purpose, the old medium was aspirated and the cells were washed with 5 mL PBS (Lonza). 1 ml of al 0.25 % Trypsin/EDTA solution (Gibco by Life

Technologies) was added and cells were incubated at 37 °C until cells detached from the culture dish. The cells were resuspended in 5 mL of pre-warmed complete medium, and pelleted by centrifugation for 5 minutes at 300 g. The pellet was resuspended in complete medium and the appropriate number of cells was transferred to a new culture dish. Medium was added to a final volume of 10 mL. The volumes that were used in this protocol were adapted to dish sizes others than 10 cm.

Cells were frozen to maintain a stock of cells at low passage number. After trypsinization, the cell pellet was resuspended in FBS containing 10 % DMSO. Aliquots of 1 mL were frozen in cryo-vials by slow cooling to -80 °C and transferred to -150 °C for long term storage. Cells were thawed for cultivation in a 37 °C water bath, resuspended in warm medium, pelleted by centrifugation, resuspended, and plated.

4.2.2 Starvation of cells from estradiol

For all experiments with controlled E_2 concentrations, cells were grown in starvation medium composed of DMEM without phenol red (Thermo Fisher Scientific), 2 % charcoal-stripped FBS (Sigma), 1 % L-Glutamine (Lonza) and 1 % Penicillin/Streptomycin (Lonza) and the desired amount of E_2 (Sigma). Medium was exchanged daily.

4.2.3 Transfection of plasmid DNA

Transient transfections of plasmid DNA into MCF-7 cells were performed with Lipofectamine® LTX Reagent with PLUS™ Reagent (Thermo Fisher Scientific) according to the manufacturer's instructions. Per 1 µg of transfected plasmid DNA, 3 µL of Lipofectamine® LTX Reagent and 1 µL of PLUS™ Reagent were used. The medium was changed 6–12 hours post-transfection. Stable transfections were carried out using FuGENE® HD Transfection Reagent (Promega) according to the manufacturer's instructions. The ratio of FuGENE® HD Transfection Reagent to transfected plasmid DNA was 3 µL per 1 µg. The medium was changed 6–12 hours post-transfection.

4.2.4 Isolation of clonal cell populations

Clonal cell populations originating from a single progenitor cell were derived by two methods: picking of colonies from a dish using cloning cylinders and sorting of single cells by fluorescence-activated cell sorting (FACS).

Cloning cylinders

After transfection of a plasmid that confers antibiotic resistance, cells were transferred to a 15 cm dish and selected with the appropriate antibiotic until all non-resistant cells died. The remaining cells were allowed to recover and form colonies. Alternatively, if the cells were not transfected and isogenic populations had to be derived from a mixed progenitor population, 300–1000 well-trypsinized cells were seeded into a 15 cm dish and cultured until colonies formed. The medium was aspirated, cells were washed with PBS, and cloning cylinders were dipped into sterile grease and placed onto the colonies to seal them from the surrounding plate. Colonies were individually trypsinized and transferred to a well of a 96-well plate to grow them for further characterization.

FACS

When transfected cells contained a fluorescent marker instead of an antibiotic resistance, individual cells were isolated by FACS. For stable cell lines, transfected cells were first isolated as a batch and grown for about two weeks to ensure stable integration of the plasmid. Cells were trypsinized and single cells were sorted through a 100 µm nozzle into a 96-well plate using an Aria III SORP cell sorter (Becton Dickinson). For better recovery, cells were sorted into medium that was supplemented with 1/3 of sterile-filtered cell culture supernatant of the same cell line (conditioned medium).

4.2.5 Generation of stable cell lines using Sleeping Beauty transposase

Transposons are mobile genetic elements that can insert themselves into DNA by a cut-and-paste mechanism. The transposase catalyzes excision of a piece of DNA that is flanked by inverted repeats and its reintegration into a random genomic position. This mechanism can be exploited to achieve efficient integration of transgenes in low-copy numbers. In order to do so, the transgene is cloned in between inverted repeats on a plasmid and transfected into the target cell together with an expression vector for the transposase. The Sleeping Beauty transposase SB100X, which was used in this study was derived by reconstituting disruptive mutations of a *Tc1/mariner*-like transposon from fish (Ivics et al. 1997) and further mutagenesis to increase its activity (Mátés et al. 2009).

The Sleeping Beauty transposase system was used to achieve a stable genomic integration of the expression cassette for GFP-labeled PP7 coat protein. MCF-7 or MCF7–noPCP_GREB1_in2_c1 cells were transfected with 3 µg of pSB-Ubc-tdPCP-tdGFP and 1.5 µg of pCMV(CAT)T7-SB100X in a 6-well dish. Cells were grown for 14 days to dilute transiently expressing cells. Single cells with a stable integration of the transgene were isolated by FACS and resulting colonies were tested for low expression levels by microscopy, yielding MCF7–SBtdPCP-tdGFP and MCF7–PCP_GREB1_in2_c1. MCF7–SBtdPCP-tdGFP cells were used for further knock-in of the PP7 sequences into exon 2 and exon 33 of *GREB1*.

4.2.6 Generation of knock-in cell lines using CRISPR/Cas9

Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas is an adaptive immune system in bacteria and archaea that utilizes an RNA-guided nuclease to specifically cleave invading genetic elements (Horvath & Barrangou 2010). Sequence specificity is achieved by Watson-Crick pairing of a guide RNA (gRNA) with a 20 bp target DNA sequence. The ease of changing this gRNA sequence in the laboratory to alter the targeting of a nuclease made this system a versatile tool for genome engineering (Cong et al. 2013, Ran et al. 2013). The system that is used most frequently is the Type II CRISPR system from *Streptococcus pyogenes* with its nuclease Cas9. After Cas9-mediated cleavage, the resulting double-strand break can be repaired by two pathways. Error-prone non-homologous end joining often results in insertion/deletion mutations and can be used to generate a gene knock-out by introducing a frame-shift in the gene of interest. Alternatively, the double-strand break can be repaired via homology-directed repair (HDR) allowing

precise modifications at the target genomic locus including introduction of custom sequences when flanked by homologous sequences to generate knock-in alleles.

To knock in the PP7 sequences together with a selection cassette into intron 2 of GREB1, a specific gRNA sequence was designed using the CRISPR design tool[2] and cloned into pX330 to yield pX330-GREB1-in2L. Homology arms for HDR, each consisting of 600–800 bp DNA around the gRNA target site, were cloned into pUC19 such that they flank the PP7 sequences and the selection cassette to yield pHR-GREB1-in2-24xPP7-LPIBCL.

MCF-7 cells were transfected with 1.5 µg of pX330-GREB1-in2L and 3 µg of pHR-GREB1-in2-24xPP7-LPIBCL in a 6-well dish and transferred to a 15 cm dish on the next day. Selection with 0.1 µg/mL Puromycin was started 3 days post-transfection and continued until colonies formed. After 3 weeks, colonies were picked using cloning cylinders and grown for further characterization. Clonal cell lines were tested by microscopy for the presence of eBFP2 labeled peroxisomes and the appearance of transcription sites after transient transfection of pHAGE-UbC-NLS-HA-tdPCP-GFP. Further confirmation of the knock-in was carried out by genotyping PCRs with primers outside of the homology arm together with primers within the knock-in cassette. This process yielded the MCF7-noPCP_GREB1_in2_c1 cell line that carries a knock-in in two *GREB1* alleles.

A similar knock-in strategy was performed for exon 2 or exon 33 of *GREB1* with pX330-GREB1-ex2T or pX330-GREB1-ex33B, respectively, instead of pX330-GREB1-in2L, and pHR-GREB1-ex2-24xPP7-LPIBCL or pHR-GREB1-ex33-24xPP7-LPIBCL, respectively, instead of pHR-GREB1-in2-24xPP7-LPIBCL to generate the knock-in cell lines MCF7-PCP_GREB1_ex2_c16 and MCF7–PCP_GREB1_ex33_c3.

4.2.7 Cre/loxP-mediated excision of selection cassette from knock-in allele

Cre recombinase of the bacteriophage P1 is a site-specific tyrosine recombinase that catalyzes the recombination between two 34 bp long recognition sites on the DNA (Abremski & Hoess 1984). These so-called loxP sites consist of two 13 bp palindromic sequences that flank an 8 bp core which provides directionality of the site. Depending on the orientation and location of the two loxP sites, the recombination can lead to insertion, deletion or inversion of DNA sequences.

The knock-in allele contains a selection cassette that is flanked by two loxP sites in the same orientation. This enables Cre recombinase-mediated excision of this selection cassette. MCF7–PCP_GREB1_ex2_c16 cells were transfected with 3 µg pCAG-Cre-IRES2-GFP in 6-well plates and BFP and GFP double positive cells with strong GFP expression were isolated by FACS four days later. After three weeks of recovery, single BFP negative cells were isolated by FACS and grown for further characterization. The excision of the selection cassette in the resulting cell line MCF7–PCP_GREB1_ex2_c16_Cre was confirmed by absence of BFP expression and by genotyping PCRs.

---

[2] Feng Zhang, Massachusetts Institute of Technology, http://crispr.mit.edu (October 2014)

4.3 Molecular biology

Standard molecular biology methods, such as preparation and transformation of chemically competent *E. coli*, amplification and isolation of plasmid DNA from bacteria, analysis of DNA on agarose gels, restriction endonuclease digestion of DNA, dephosphorylation of DNA using alkaline phosphatase, ligation of DNA fragments with T4 DNA Ligase, assembly of DNA fragments using In-Fusion cloning, purification of DNA from agarose gels as well as PCR reactions, and spectrophotometric quantification of DNA were carried out according to manufacturer's instructions of the kits and enzymes listed in 4.1.3 and 4.1.4, and standard protocols (Green & Sambrook 2012).

4.3.1 Polymerase chain reaction (PCR)

Cloning PCRs

All PCRs that were performed to amplify DNA for further use in cloning protocols were performed with the Q5® High-Fidelity DNA Polymerase (New England Biolabs) and primers from Table 6. The reaction was set up according to Table 13 and the temperature program in Table 14 was used with annealing temperatures according to the primer pair.

Table 13: Setup of Q5 cloning PCR reaction.

Reagent	Concentration	Amount
Q5 buffer	5 x	5 µL
template DNA	variable	500 pg (plasmid) or 100 ng (genomic DNA)
forward primer	10 µM	1.25 µL
reverse primer	10 µM	1.25 µL
dNTPs	10 mM each	0.5 µL
Q5 Polymerase	2 U/µL	0.5 µL
ddH_2O	55.5 M	to 25 µL

Table 14: Temperature program for Q5 PCR reaction.

Temperature	Time	
98 °C	30 sec	
98 °C	10 sec	25–35 cycles
65–72 °C	10 sec	
72 °C	30 sec / kb	
72 °C	2 min	
4 °C	forever	

Genotyping PCRs

Genotyping PCRs were performed on genomic DNA to check for correct insertion of the transgene and correct recombinase-mediated excision of the selection cassette from the transgene. PCRs were set up according to Table 15 using the TrueStart Hot Start *Taq* DNA Polymerase (Thermo Fisher Scientific) and primers from Table 7. A touchdown PCR was performed according to Table 16 to enhance specificity.

Table 15: Setup of genotyping PCR reaction.

Reagent	Concentration	Amount
TrueStart buffer	10 x	2 µL
genomic DNA	variable	100 ng
forward primer	10 µM	0.38 µL
reverse primer	10 µM	0.38 µL
$MgCl_2$	25 mM	1.6 µL
dNTPs	2 mM each	2 µL
TrueStart Polymerase	5 U/µL	0.16 µL
ddH_2O	55.5 M	to 20 µL

Table 16: Temperature program for genotyping PCR reaction.

Temperature	Time	
95 °C	4 min	
95 °C	30 sec	
60 °C x 2, 57 °C x 2, 54 °C x 4, 52 °C x 4, 50 °C x 26	30 sec	in total 38 cycles
72 °C	70 sec	
72 °C	10 min	
4 °C	forever	

4.3.2 Isolation of genomic DNA from mammalian cells

Total genomic DNA was isolated from cells using the DNAzol® Reagent (Thermo Fisher Scientific) according to manufacturer's recommendations. The method uses cell lysis in a mixture of guanidine thiocyanate and detergents followed by ethanol precipitation (Chomczynski et al. 1997). The DNA was directly used in cloning or genotyping PCRs.

4.3.3 Isolation of total RNA from mammalian cells

Total RNA was isolated using TRIzol® Reagent (Thermo Fisher Scientific) according to manufacturer's recommendations. To isolate RNA, cells were homogenized in TRIzol, chloroform was added and RNA was precipitated from the aqueous phase using isopropanol (Chomczynski 1993).

For RT-qPCR experiments, 2×10^5 cells were seeded into a well of a 6-well plate and grown in starvation medium for 3 days. Transcription was induced overnight (~18 hours) with different E_2 concentrations to record the E_2 dose-response curve. Alternatively, to measure kinetics of RNA induction, cells were induced with either 10 pM or 1000 pM E_2 and samples were collected every 10 minutes for 2 hours. RNA extraction was carried out in 600 µl of TRIzol according to manufacturer's instructions. The RNA was used for cDNA generation and quantification of gene expression by RT-qPCR.

4.3.4 Quantification of gene expression by RT-qPCR

Generation of cDNA by reverse transcription (RT)

Reverse transcription of RNA into cDNA was performed using the SuperScript® II RT (Thermo Fisher Scientific) with random hexamers. The reaction was set up according to Table 17, incubated for 5 minutes at 65 °C and placed on ice. Then, the components from Table 18 were added and the mix was incubated for 10 minutes at 25 °C and for 50 minutes at 42 °C. Finally, heat inactivation was performed for 15 minutes at 70 °C.

Table 17: Setup of RT reaction (first mix).

Reagent	Concentration	Volume
RNA	variable	variable (400 ng)
random hexamers	50 µM	1 µL
dNTPs	10 mM each	1 µL
ddH_2O	55.5 M	to 12 µL

Table 18: Setup of RT reaction (second mix).

Reagent	Concentration	Volume
SuperScript buffer	5 x	4 µL
DTT	0.1 M	2 µL
RNAse OUT	40 U/µL	1 µL
SuperScript II RT	200 U/µL	1 µL

Quantitative PCR (qPCR)

Quantitative PCRs were performed on 1 µL of template cDNA, genomic DNA or diluted plasmid DNA with primer pairs listed in Table 8. The efficiency of amplification was determined for all primer pairs using serial dilutions of template DNA and confirmed to be above 90 %. All measurements were performed as technical duplicates in a ViiA™ 7 Real-Time PCR System (Thermo Fisher Scientific) in a 384-well format (Roche).

Table 19: Mix for a single qPCR reaction.

Reagent	Concentration	Volume
cDNA	variable	1 µL
forward primer	10 µM	0.1 µL
reverse primer	10 µM	0.1 µL
Power SYBR Green	2 x	5 µL
ddH_2O	55.5 M	4 µL

Table 20: Temperature program for qPCR.

Temperature	Time	
50 °C	2 min	
95 °C	10 min	
95 °C	15 sec	40 cycles
60 °C	1 min	
60–95 °C		melt curve
4 °C	forever	

Relative quantification of gene expression

Threshold cycles (Ct) were calculated using a manually set detection threshold to allow comparability between plates. Mean Ct-values of technical replicates were used. To compare gene expression between samples, normalization to the reference gene *GAPDH* was performed. The Ct-value for *GAPDH* was subtracted from the Ct-value of the gene of interest to obtain the ∆Ct value. This value was used to calculate expression levels as percent of *GAPDH*:

$$\% \; GAPDH = \; 2^{-\Delta Ct} \times 100 \; \% \qquad \textbf{(1)}$$

The difference of ΔCt values, the ΔΔCt value, was calculated between samples and used to calculate the relative difference in expression:

$$\text{change in expression} = 2^{\Delta\Delta Ct} \quad \textbf{(2)}$$

When calculating fold changes, for example to time point 0, all values were divided by the mean expression value from biological replicates at that reference point. Standard deviations were calculated from biological replicates, in case of fold changes after normalization to the reference.

In case of allele-specific RT-qPCR, when expression levels of the wildtype and knock-in allele were compared, extreme care was taken to correct for amplification bias due to primer efficiency. Therefore, normalization was carried out using an external plasmid standard that contained all PCR products. qPCR was performed on a serial dilution of the plasmid and Ct-values were plotted against the $\log_{10}$ of ng plasmid. Slope and intercept of a linear regression were calculated and used to determine the absolute amount of target DNA in the cDNA reaction. The calibrated expression levels were calculated as percent of *GAPDH* using equation 3.

$$\%\ GAPDH = 10^{\frac{Ct_{target} - intercept_{target}}{slope_{target}}} \Big/ 10^{\frac{Ct_{GAPDH} - intercept_{GAPDH}}{slope_{GAPDH}}} \times 100\% \quad \textbf{(3)}$$

4.4 Live-cell imaging of nascent transcription

4.4.1 Preparation of cells

Cells were trypsinized three days prior to imaging and 30 µL of a 0.6 × 10^6 cells/mL solution were filled into each channel of a 6 channel µ-Slide VI 0.4 ibiTreat (ibidi) and 120 µl of medium was added to fill the reservoirs. The medium was replaced by starvation medium containing the desired amount of E_2 (Sigma) on the next day and changed daily. To observe the effect of small molecule inhibitors, 2.5 mM sodium butyrate (Sigma), 1 µM PFI-1 (Sigma), 10 µM C646 (Sigma), or DMSO (Sigma) (final DMSO concentration in all samples 0.05 %) were added four hours prior to imaging and after growing cells in 20 pM E_2 for two days. For induction experiments, cells were grown in starvation medium without E_2 for at least 48 hours and placed into the microscope. After 51 minutes of imaging the medium was replaced with starvation medium containing either 10 pM or 1000 pM E_2.

4.4.2 Image acquisition

Live-cell images were acquired on a DeltaVision™ Elite microscope system (GE Healthcare) equipped with an environmental control chamber (Imsol) and a CO_2 mixer (Leica) to maintain 37 °C and 5 % CO_2 during imaging experiments. Excitation light was generated using a 7 color InsightSSI module and focused through a 60× 1.42 NA Oil Plan APO objective. Excitation and collection of emitted fluorescence of eGFP was achieved using the FITC filters and the polychroic beam splitter for DAPI, FITC, TRITC and Cy5. Images were acquired on a pco.edge sCMOS-camera operating in 2x2 binning mode,

yielding an image with 1024 × 1024 pixels and a pixel size of 216 nm. The microscope was controlled via softWoRx v6.5.2.

To image transcription in living cells, z-stacks with 12 planes spaced 0.55 µm apart were acquired every 3 minutes for 260 time points (total imaging time ~13 hours) in the FITC channel with 2 % light intensity with 100–120 ms exposure times. One brightfield image (POL channel) was acquired with an exposure time of 50 ms at 5 % light intensity at each time point to follow cell viability. The first 10 frames of each movie were discarded to remove initial photobleaching of the medium. Photobleaching during the rest of the movie was not corrected for. For induction experiments, images were acquired with the same settings every 1.5 minutes for 200 time points (300 x 5 minutes for analysis of daughter cells) without discarding initial images. Images for visualization of single transcripts were acquired as a z-stack with 14 slices spaced 0.27 µm apart, with 100 % light intensity and 100–120 ms exposure time.

Imaging of the E_2 dose-response was performed for the 0–20 pM E_2 datasets simultaneously on the same day and for 100–1000 pM E_2 datasets on a different day. Data for E_2 induction is a combination of two separate experiments for each E_2 concentration. The dual allele dataset and the inhibitor dataset are from one experiment each.

4.4.3 Live-cell image analysis

All live-cell image analysis was performed by custom MATLAB scripts.

Segmentation of nuclei

Nuclei were labeled by nuclear localized tdPCP-tdGFP. Accordingly, nuclear segmentation was directly carried out on mean intensity projections of images that were acquired in the GFP channel. In a first step, bright nuclear foci that adversely influence segmentation were removed by setting an intensity cut-off at the 92nd percentile on the local background (Gaussian smoothing with width of 30 µm) subtracted image. The remaining image was scaled to intensity values between 0 and 1, smoothened by applying a Gaussian filter (width of 0.8 µm) and the local background was subtracted before a user selected threshold (usually between 0.03 and 0.06) was applied. Holes in the resulting mask were filled and the mask was smoothened by applying an opening operation with a disk structuring element with a radius of 2.5 µm. Nuclei that were in close proximity and could not be separated by this approach were identified by size (area > 260 µm²) and a shape measure (solidity < 0.93). These clustered nuclei were then iteratively separated by identifying the best watershed lines that connect two concave regions in a nuclear mask such that they lead to a separation of clustered nuclei (described in Stoeger et al. 2015)[3]. Finally, objects that were smaller than 60 µm² or bigger than 1300 µm² were removed to yield the final nuclear mask.

[3] Lucas Pelkmans, https://github.com/pelkmanslab/ImageBasedTranscriptomics (January 2016)

Tracking of nuclei

Nuclei were identified as described above in each frame of a time-lapse movie. Furthermore, to generate a complete set of corresponding masks for each nucleus at each time point, the result of the nuclear segmentation of the previous time point was used to correct possible errors in the segmentation of the current time point. To detect possible errors such as over- or undersegmentation of nuclei, as well as the disappearance or the appearance of new nuclei over time, the pixel-based overlap of all individual nuclear masks of both time points were calculated. Because cells move slowly with respect to the imaging interval, nuclear masks of corresponding cells should show high overlap in successive images. Six cases of frame-to-frame correspondences were distinguished:

1. One-to-one: Only the mask of the previous and the current frame showed overlap with each other indicating that those nuclei correspond to each other.
2. One-to-zero: A mask from the previous frame did not show any overlap with a mask from the current frame. The nucleus was lost. To avoid a loss of a nucleus that was sometimes identified as a false-negative, an image registration algorithm was used to identify the best transformation between the previous nuclear image and the image at the same position in the current frame and yielded a new position for the mask. Image registration was performed for a maximum of five consecutive frames before the nucleus was deleted.
3. Zero-to-one: A mask in the current frame did not overlap with any mask in the previous frame. A new nucleus was created.
4. Many-to-one: Two or more masks of the previous frame overlapped with a single mask in the current frame as a result of undersegmenting the nucleus in the current frame. To resolve this undersegmentation, the points along the outline of the previous masks were aligned with the points along the outline of the current mask using an iterative closest point (ICP) algorithm (Bergström & Edlund 2014)[4] and the points of the current mask were assigned to the nucleus that the closest point of the previous outlines belonged to. New masks were created from the outlines.
5. One-to-many: A single nuclear mask of the previous frame overlapped with two or more masks of the current frame as a result of oversegmentation in the current frame. To resolve this oversegmentation, a similar ICP-based approach as above was applied to fuse the masks.
6. Many-to-many: Many nuclear masks of the previous frame overlapped with many masks of the current frame. This most complex case could not be fully resolved and often led to erroneous assigned correspondences.

After finding frame-to-frame correspondences in the corrected masks, incompletely tracked nuclei, such as lost or newly appeared nuclei, were discarded. Thereby, dividing and dying cells were removed. Nuclei that touched image borders or that showed errone-

[4] Per Bergström, http://www.mathworks.com/matlabcentral/fileexchange/12627-iterative-closest-point-method (September 2015)

ous tracking were manually removed. In a typical movie with a cell density of 80 % about 30 nuclei could be completely tracked.

Bandpass filtering

A two-dimensional bandpass filter was used to reduce pixel noise and background fluorescence, e.g. of unbound nuclear tdPCP-tdGFP. Edges of the image were replicated, the image was transformed into Fourier space and multiplied with a filter that was calculated as follows.

$$BP_{x,y} = \frac{1}{1 + \left(\frac{2d_{x,y}C_{HP}}{w+h}\right)^{12}} - \frac{1}{1 + \left(\frac{2d_{x,y}C_{LP}}{w+h}\right)^{4}} \quad \textbf{(4)}$$

with w and h being the width and height of the filter, respectively, $d_{x,y}$ being the distance of the pixel at position x/y to the center of the image, and C_{HP} and C_{LP} are the cutoffs (in px) for high pass and low pass, respectively.

Spot detection

Spots were detected on maximum intensity projections of bandpass-filtered (0.65 µm to 4.3 µm) images. Spot detection was performed using the u-track package for MATLAB in version 2.0 (Jaqaman et al. 2008) with a user-defined width of the point-spread-function (0.4 µm) and an alpha-value of 0.13.

Tracking of transcription sites

Transcription sites were relatively immobile in the nucleus and showed only minimal movement relative to the nucleus over the timeframe of imaging experiments. This property was used to track the transcription site within a nucleus and follow its position even in the absence of a visible spot. Movement of the nucleus consists of translation and rotation, both of which were inferred from the nuclear outlines.

The outlines of a nucleus from two consecutive time points were superimposed and aligned via an ICP algorithm to infer their transformation. Big changes in the size of the nucleus, which resulted from detection errors, adversely affected the alignment. In such cases, the ICP algorithm was repeated 10 times with shifted 66 % contiguous points of the outline of the bigger nucleus and the best transformation was used.

After inference of nuclear transformations, the positions of all spots that were not more than 1 µm away from the nucleus were transformed such that they were fixed relative to the given nucleus. Then, tracklets were generated using the u-track package for MATLAB using a linear motion Kalman filter with a search radius of 6 without merging or splitting and without gap closing. These short and incomplete tracklets only cover a fraction of all time points of a time-lapse movie and need to be linked and interpolated to generate complete tracks that last from the beginning to the end of a movie. Only tracklets that cover at least four time points were considered in the tracklet linking step. A brute-force strategy was chosen, such that all possible combinations of the up to 18 longest tracklets were considered and a cost for the linkage was calculated according to equation 5.

$$cost = \begin{cases} \infty; & \text{if tracklets overlap in time} \\ \sum_{t=1}^{260} \frac{10 i_t}{i_{max}} + n_{gaps} + \sum_{all\ linkages} \frac{2d^2}{\Delta_t}; & \text{otherwise} \end{cases} \quad \textbf{(5)}$$

i_t … u-track spot intensity in frame t
i_{max} … 90 % percentile of all u-track spot intensities in nucleus
n_{gaps} … Number of frames without tracklet
d … Distance between end and start of consecutive tracklets (in px)
Δ_t … Number of frames without tracklet for this linkage

The combination of tracklets with the lowest cost was chosen and the gaps in between tracklets were filled using linear interpolation before the positions were retransformed into the original coordinates of the movie. A second position in each nucleus was used to describe the background that is generated by the spot quantification algorithm. This position was set to the centroid of the nuclear mask and shifted such that it is at least 3 µm away from the position of the transcription site.

All generated tracks were reviewed, and erroneous assigned positions were corrected manually. Errors were mainly observed in case of dim transcription sites that disappear for a long time in combination with strong rotational movement of a nucleus. Cells where the transcription site moves out of focus or divides during acquisition were discarded.

Quantification of fluorescence intensities

Fluorescence intensities were quantified on bandpass filtered images (0.4 – 4 µm) in which pixel noise and the background of unbound tdPCP-tdGFP is reduced. The intensity was quantified by fitting a 3-dimensional Gaussian distribution with an offset (see equation 6) to the intensity distribution in the 3D image stack in a circular window centered around the tracked position that had a height and width of 1.9 µm. The squared error was minimized using the fminsearchbnd function[5] with σ_{xy} being constrained to 0.1 – 0.4 µm and σ_z to 0.3 – 1 µm.

$$I_{xyz} = c + A\, e^{-\frac{1}{2}\left(\left(\frac{x-x_0}{\sigma_{xy}}\right)^2 + \left(\frac{y-y_0}{\sigma_{xy}}\right)^2 + \left(\frac{z-z_0}{\sigma_z}\right)^2\right)} \quad \textbf{(6)}$$

I_{xyz} … Intensity at position x, y, z
A … Amplitude
x_0, y_0, z_0 … Center of Gaussian distribution
σ_{xy}, σ_z … Width of Gaussian distribution in xy and z
c … Offset

[5] John D'Errico, https://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd--fminsearchcon (February 2014)

The integrated intensity was then calculated according to equation 7.

$$I = A\,{\sigma_{xy}}^2\,\sigma_z\,(2\pi)^{\frac{2}{3}}\,\Delta_z \quad \textbf{(7)}$$

I … Integrated intensity
Δ_z … Distance between z-slices

Absolute quantification by calibration to intensities of single RNAs

To generate absolute numbers of transcripts from the arbitrary number generated by the fluorescence quantification, it was necessary to calibrate the intensities to a known standard. When cells at saturating concentrations of E_2 were imaged at maximum excitation settings (100 % instead of 2 % light intensity), low intensity spots were visible in the vicinity of transcription sites. These spots are likely single finished transcripts that have left the site of active transcription and are diffusing within the nucleoplasm. They were used to quantify the fluorescence intensity of a single transcript, which was then used as a normalization factor to derive absolute transcript numbers.

Dim spots that likely represent single transcripts were manually identified in the images and quantified as described above. A Gaussian distribution was fitted to the resulting intensity histogram and the mean of the distribution was used as the mean intensity for a single transcript. To adjust for the relative difference in illumination between 2 % and 100 % light intensity, images were acquired with both illumination conditions at the same position and the intensities of transcription sites were quantified. A linear function was fitted to the ratio of intensities and their slope was used as a normalization factor. The procedure was repeated every time the setup of the microscope or image acquisition parameters changed.

Additionally, histograms of spot intensities were matched with spot intensities from smRNA FISH, where absolute RNA numbers are known (see 4.5.4 and 4.8.2).

Feature extraction from time traces

Raw fluorescence intensities were converted in absolute RNA numbers by division by the intensity of a single transcript. A running median with a window size of five time points was applied to smooth the time trace. The slope of the curve was calculated as the difference between two consecutive time points. This slope was subsampled 10-fold by linear interpolation and a threshold of 0.65 transcripts per 3 minutes was applied. Gaps and peaks with duration of less than one imaging interval were discarded. ON- and OFF-times were derived from the time the slope is above or below the threshold, respectively. The burst size was calculated for each ON-period as the difference of intensity of the smoothed time trace. The initiation rate was calculated for each burst by dividing the burst size by the ON-time. ON- and OFF-times were also calculated for regions that encompass beginning or end of the time trace, to include long OFF-times for non-responders at low E_2 concentrations. The area under the curve (AUC) was calculated as the sum over all intensities. To calculate response-times from induction experiments and simulations, trajecto-

ries were median filtered with a window of seven time points. Then the time from which on the trajectory stayed above a threshold of two transcripts for at least five consecutive time points was determined.

4.5 Single-molecule RNA fluorescence *in-situ* hybridization

4.5.1 Preparation and fixation of cells

Cells were seeded into µ-Slide VI 0.4 ibiTreat slides and grown with different E_2 concentrations (with or without 1 µM ICI 182,780 as indicated) as described in section 4.4.1. After three days, cells were washed with PBS and fixed with 4 % PFA in PBS for 10 minutes at room temperature. Cells were washed twice with PBS and permeabilized with 70 % ethanol at 4 °C for at least 1 hour.

4.5.2 Probe hybridization and mounting

Custom Stellaris® FISH Probes were designed against introns 2, 9 and 10 of human *GREB1* by utilizing the Stellaris® RNA FISH Probe Designer (version 4.2) and obtained from LGC Biosearch Technologies labeled with Quasar® 670. Stellaris® FISH Probes recognizing exon 5–9 in human *GREB1* were obtained labeled with Quasar® 570 (Catalog #VSMF-2158-5). The hybridization with both probe sets followed the manufacturer's instructions with slight modifications. Briefly, permeabilized cells were rehydrated for 5 minutes with wash buffer A and the buffer was removed from the channels prior to adding 50 µL hybridization solution. Hybridization was performed overnight at 37 °C in the dark.

Table 21: Composition of wash buffer A.

Reagent	Concentration	Amount
Stellaris® RNA FISH wash buffer A	5 x	1 mL
Deionized formamide	100 %	500 µL
Nuclease-free H_2O	55.5 M	3.5 mL

Table 22: Composition of smRNA FISH hybridization solution.

Reagent	Concentration	Amount
Stellaris® RNA FISH hybridization buffer	-	270 µL
Deionized formamide	100 %	30 µL
Intronic *GREB1* probes (Quasar® 670)	12.5 µM	3 µL
Exonic *GREB1* probes (Quasar® 570)	12.5 µM	3 µL

Samples were washed twice with wash buffer A and once with wash buffer A containing 5 ng/µl DAPI, each for 20 minutes at 37 °C. Cells were washed in Stellaris® RNA FISH wash buffer B for 5 minutes at room temperature and equilibrated for 2 minutes in GLOX buffer. The buffer was removed, the channels were filled with 50 µL GLOX anti-fade solution, and the samples were imaged immediately.

Table 23: Composition of GLOX buffer.

Reagent	Concentration	Amount
Tris HCl, pH 7.5	1 M	100 µL
20 x SSC (150 mM NaCl, 15 mM sodium citrate, pH 7)	20 x	1 mL
glucose	20 %	200 µL
Nuclease-free H_2O	55.5 M	to 10 mL

Table 24: Composition of GLOX anti-fade solution.

Reagent	Concentration	Amount
GLOX buffer		300 µL
Trolox	1 mM	3 µL
glucose oxidase	3.7 mg/mL	3 µL
catalase	2 U/µL	3 µL

4.5.3 Image acquisition

Images were acquired on a DeltaVision™ Elite microscope system with a 60× 1.42 NA Oil Plan APO objective. Excitation and collection of emitted fluorescence was achieved using the filters and the polychroic beam splitter for DAPI, FITC, TRITC and Cy5. Images were acquired on a pco.edge sCMOS-camera without binning, yielding an image with 2048 × 2048 pixels and a pixel size of 108 nm. The microscope was controlled via softWoRx v6.5.2. z-stacks with 32 planes spaced 0.27 µm apart were acquired in the TRITC channel (50 % light intensity, 200 ms exposure) for Quasar® 570, the Cy-5 channel (32 % light intensity, 100 ms exposure) for Quasar® 670, the FITC channel (100 % light intensity, 100 ms exposure) for GFP, and the DAPI channel (50 % light intensity, 100 ms exposure). One brightfield image was acquired with an exposure time of 50 ms at 5 % light intensity.

4.5.4 Single-molecule RNA FISH image analysis

All image analysis was performed by custom MATLAB scripts.

Segmentation of nuclei

Nuclei were detected on maximum intensity projections of images in the DAPI channel. Intensity inhomogeneity was removed by local background (Gaussian smoothing, width of 10 µm) subtraction. After scaling of intensities to range from 0 to 1, global thresholding using Otsu's method (Otsu 1979) was applied. Holes in the mask were filled and the mask was smoothened by applying an opening operation with a disk structuring element with a radius of 2.5 µm. Clustered nuclei were identified (area > 175 µm², solidity < 0.935) and separated as described in 4.4.3. Nuclei that were smaller than 80 µm² or bigger than 250 µm² were removed and the remaining objects were inflated by 4 px. Nuclei with a mean DAPI intensity of more than 3000 and a solidity of less than 0.96 were not considered for further analysis.

Spot detection and quantification

Spots were detected on maximum intensity projections of bandpass-filtered (0.4 µm to 4 µm) images. Spot detection was performed using the u-track package for MATLAB in version 2.0 (Jaqaman et al. 2008) with a point-spread-function width of 0.12 µm and an

alpha-value of 0.1. Spots with an amplitude above 0.004 (Quasar® 570 and GFP) or 0.0015 (Quasar® 670) were quantified by fitting a three-dimensional Gaussian function (see equation 6) to the 3D image stack within a circular window of 1.4 µm with σ_{xy} being constrained to 0.08 – 0.38 µm and σ_z to 0.27 – 1.35 µm. The integrated intensity was then calculated according to equation 7. Spurious spots (σ_{xy} < 0.12 µm, amplitude < 300 (Quasar® 570) or < 100 (Quasar® 670 and GFP), ratio of amplitude and (σ_{xy}–1) < 800 (Quasar® 570), < 200 (Quasar® 670), or < 150 (GFP) were discarded.

Absolute quantification by calibration to intensities of single RNAs
Single transcripts are visible as diffraction-limited spots in the cytoplasm of cells when smRNA FISH for exons is performed. The intensity of these spots was used to calibrate the intensities of transcription sites and derive absolute nascent RNA numbers. A lognormal distribution was fitted using the lsqnonlin function in MATLAB to the intensities of spots that are located outside of nuclei to yield the mean intensity of a single transcript.

4.6 High-content imaging of transcription

4.6.1 Preparation of cells

10^4 cells were seeded per well of a 96-well SensoPlate™ Plus glass bottom microplate (greiner) three days prior to image acquisition. The medium was replaced with starvation medium containing the desired concentration of E_2 every day and DMSO and small molecule inhibitors were added four hours prior to fixation. Cells were washed with PBS and fixed with 4 % PFA in PBS for 10 minutes on ice. Afterwards, cells were washed twice with PBS and a nuclear counter-staining was performed with 0.5 µg/mL DAPI (Sigma) in PBS for five minutes or 2.5 µM DRAQ5 (eBioscience) for 30 minutes. Cells were stored in PBS at 4 °C until image acquisition. Cell culture and treatment for high-content imaging experiments was performed by Daria Steinshorn.

4.6.2 Image acquisition

Seven fields were imaged per well in an Opera Phenix™ High Content Screening System (PerkinElmer) with a 20× 1.0 NA water immersion objective using spinning disc confocal mode. A z-stack with 22 planes spaced 1.2 µm apart was acquired without binning, resulting in a pixel size of 0.30 µm. Exposure time in the EGFP channel (excitation: 488 nm laser, emission: 500–550 nm) was 500 ms at 100 % illumination intensity. Nuclei were imaged either in the DAPI channel (excitation: 405 nm laser, emission: 435–480 nm) for 60 ms at 80 % intensity or in the DRAQ5 channel (excitation: 640 nm laser, emission: 650–760 nm) for 300 ms at 50 % intensity depending on the nuclear counterstain.

High-content imaging for the E_2 dose-response was performed in three separate experiments with technical duplicates. Inhibitor treatment was performed twice with technical duplicates.

4.6.3 High-content image analysis

Images from the high content screening microscope were analyzed using the Harmony® High Content Imaging and Analysis Software (PerkinElmer). First, maximum intensity projections were calculated for each channel and a combined DAPI-EGFP image was calculated by summing the EGFP channel and 1/10 of the intensity of the DAPI channel (for DRAQ5: EGFP + 1/4 × DRAQ5). This image was used in the “find nuclei” building block to identify nuclei using method C (parameters: common threshold 0.45 (0.7 for DRAQ5), area 30 µm, split factor 20.5 (10 for DRAQ5), individual threshold 0.45, contrast 0.05). Nuclei that touch image borders were removed and good nuclei were selected based on intensity and morphology properties (mean GFP intensity < 300, nucleus roundness > 0.65, nucleus area < 300 µm², nucleus ratio width to length > 0.45, nucleus area > 80 µm). The plane from which the maximum pixel intensity originated during the maximum intensity projection was used to discard out-of-focus nuclei. Nuclei were removed if the mean plane map value was below 4 or above 18. Spots were identified in a region that encompasses the nucleus plus a rim of 1 µm on a background subtracted image (EGFP channel - 5 px Gaussian filtered EGFP channel) using method C (parameters: radius ≤ 1.61 µm, contrast > 0.82 µm, uncorrected spot to region intensity > 2.5, distance ≥ 3 µm, spot peak radius 0.24 µm). A linear classifier was used to discriminate real transcription sites from spurious spot detections (discarded if 14.4 × spot contrast + 0.0324 × spot area + 0.126 × spot background intensity + 0.158 × spot to region intensity < 17.06). Spot and nuclei properties were exported and analyzed using custom scripts in R. Spot intensities were calculated as the product of the corrected spot intensity and the spot area. Only the brightest spot was considered for each nucleus.

Spot localization was analyzed with a custom building block. The localization is expressed as the distance between the center and the edge of the nucleus as a value between 0 (edge) to 1 (center). A random distribution was approximated as follows. First, a number in the interval between 0 and (radial scale + 1 µm – spot radius)² was sampled for each nucleus. Then the square root of this value was divided by the radial scale and subtracted from 1.

4.7 Bayesian inference on single-cell transcription time traces

Implementation of stochastic simulations and the Sequential Monte-Carlo Approximate Bayesian Computation algorithm, as well as all model fitting to experimental data was performed by Stephan Baumgärtner. Implementation details can be found in his PhD thesis (Baumgärtner 2017) and the complete implementation is publicly available on Github[6]. The general procedure is outlined below.

[6] https://github.com/baumgast/gene_transcription_SMC_ABC

4.7.1 Stochastic modeling of promoter progression and transcription

Promoter progression was modeled as an abstract cycle of promoter states, where the promoter stochastically switches from one state into the next in an irreversible, ratchet-like manner. Each of these steps can be thought of as a series of protein binding events that leaves a stable modification of the chromatin template. Each of the states is either transcriptionally active (ON) or inactive (OFF) and switching between them occurs with rates k_{ON} and k_{OFF}, respectively. The simplest model is a two-state model with a single ON- and a single OFF-state. More complex model topologies were considered by adding additional states either in the OFF- or the ON-phase up to a model with 10 states (see Table 25). The additional states in bigger models lead to more free parameters.

Table 25: Topologies and number of parameters for cyclic promoter models. The transition rates between states plus the initiation rate are approximated by fitting, while the elongation rate is fixed. Parameter numbers are given for a model without extrinsic noise. Additional free parameters can occur depending on the extrinsic noise source (Table 26).

Name	ON-states	OFF-states	Free parameters
1-1	1	1	3
1-2	1	2	4
2-2	2	2	5
1-9	1	9	11
2-8	2	8	11

Transcripts could be initiated within an ON-state with rate k_{init}. Stochastic simulations of promoter progression and initiation were performed using the stochastic simulation algorithm (Gillespie 1977). While other models describe transcription of RNA as instantaneous, here the kinetics of polymerase elongation was considered and produced a characteristic profile of fluorescence intensity, depending on the gene structure and the position of the PP7 stem-loops. For *GREB1*, the stem-loops were modeled to occur 3 kb after transcriptional initiation and the transcript length is 88 kb. Depending on the rate of polymerase elongation, k_{elong}, the fluorescence profile observed for a single *GREB1* transcript changes. Elongation was assumed deterministic, such that a fluorescence profile of one RNA was added at the sampled time of each initiation event. The summed-up signal of all RNAs was resampled with the imaging time interval and noise was added to produce the final simulated fluorescence trajectory. Noise was assumed to be log-normal distributed with mean and width that were estimated from fluorescence quantifications at random positions in the nucleus.

4.7.2 Implementation of extrinsic noise

Cell-to-cell variability was implemented by resampling of parameters prior to each single-cell simulation. Such resampling assumes that extrinsic variability is stable over time. Furthermore, it was assumed that extrinsic noise affects all *GREB1* alleles similarly during dual-allele simulations. Eight different possible noise sources were considered: Each kinetic parameter (k_{ON}, k_{OFF}, k_{elong}, and k_{init}) alone, as well as in combination with k_{elong} was resampled, plus a model without noise. The width of the distribution for resampling was added as a parameter to the model (Table 26).

Table 26: Parameters of extrinsic noise sources. σ denotes the scale parameter of the normal distribution used for resampling.

Index	Resampled parameter(s)	Sampling distribution	Additional parameter
0	none	-	-
1	k_{elong}	uniform(1,5) [kb/min]	-
2	k_{init}	normal(k_{init} ,σ_{init}) [min^{-1}]	σ_{init}
3	k_{ON}	normal(k_{ON} ,σ_{ON}) [min^{-1}]	σ_{ON}
4	k_{OFF}	normal(k_{OFF} ,σ_{OFF}) [min^{-1}]	σ_{OFF}
5	k_{elong} and k_{init}	combination of above	σ_{init}
6	k_{elong} and k_{ON}	combination of above	σ_{ON}
7	k_{elong} and k_{OFF}	combination of above	σ_{OFF}

The combination with five possible promoter topologies lead to 40 different models that were evaluated during model selection.

4.7.3 Sequential Monte-Carlo Approximate Bayesian Computation

A Bayesian method for model fitting

The model that was used in this thesis consists of different topologies and incorporates polymerase elongation to approximate the signal-generating process during transcription. It was not possible to calculate the likelihood for a given set of parameters and topologies for such a process. A likelihood-free Bayesian method, called Approximate Bayesian Computation, was used, which is able to estimate parameters and select between model topologies at the same time (Beaumont et al. 2002). It relies on comparisons of multiple simulations with the data using a distance metric, thereby eliminating the need for likelihood calculations. A sequential Monte-Carlo version of ABC was used that iteratively refines a population of particles, in which each particle consists of parameters and a model topology. The final particle population yields the approximate posterior distributions for parameters and model topologies.

Distance metric to compare simulations and experimental data

Five different features were calculated for each simulated dataset and compared to the same features of the experimental data to assess how well both datasets resemble each other. The sum of all values was used as the distance function during SMC ABC, with smaller values indicating better agreement of simulations with data.

1) *Global distribution of intensity values.* The distribution of all measured intensities characterizes the amount of time spent transcribing and the intensity of transcription sites but ignores the time-course nature and differences between individual cells. Distributions of data and simulations were compared using the Kolmogorov-Smirnov statistic.

2) *Mean autocorrelation function.* The autocorrelation function (ACF) describes the correlation of intensities when shifted by various time lags. It captures for how long the signal is similar to itself. The ACF was calculated as the average over 25 sliding windows of 125 time points each. The distance between two datasets was calculated as the sum of squared distance.

3) *Distribution of ACF half-lives*. The half-live of the ACF was calculated for each fluorescence trajectory as the mean time of crossing the value 0.5 over all window positions. The distribution of this value captures cell-to-cell variability and the similarity of data and simulations was compared using the Kolmogorov-Smirnov statistic.

4) *Distribution of ACF values at a lag of one.* The value of the ACF at the first lag partitions responding and non-responding cells. Its distribution was compared between data and simulations using the Kolmogorov-Smirnov statistic.

5) *Maximum mean discrepancy of all intensities.* This statistic compares multivariate distributions (Gretton et al. 2012).

The stochastic nature of the simulations leads to a distance value greater than zero even when parameters for simulations are the same. A value of 0.5 was estimated for an optimal fit from repeated simulations with the same set of parameters.

Generation of a start population
A start population of 50.000 particles was generated using prior beliefs about model topology and parameters. Simulations were performed for each particle and compared to experimental data as described above. The best 2000 particles were chosen as start population for further iterations.

Iterative refinement of parameters
In each iteration, the best 20 % of particles of the previous iteration were used to generate new particles. Parameters and topologies of new particles were sampled in the proximity of old values. The new particles were simulated and accepted if the distance was better than the worst of the initial 20 %. By iterative refinement of parameters, posterior distributions are approximated. The algorithm was terminated when the improvement in the distance measure of the worst particle was less than 2 %.

4.7.4 Benchmarking

The ability of the SMC ABC algorithm to recover model topology and values of kinetic parameters was analyzed using synthetic datasets. They were generated by forward simulations with known parameters and then fitted. All kinetic parameters were varied, and excellent recovery of values was observed (Figure 28). However, very short OFF-times (less than 15 minutes) were hardly recovered and the contribution of ON-time and initiation rate to the burst size was not separable. Model selection performed well, with the tendency to retrieve related extrinsic noise sources.

4.7.5 Global model fitting

In order to evaluate whether differences between datasets can be explained by changing only a single kinetic parameter, multiple datasets from various experimental conditions were globally fitted at once. This reduced the number of free parameters during fitting and reduced the risk of overfitting. The posterior distributions of the individual fits were used to extract the most frequently chosen model topology and fixed this during global fitting. For “global” parameters, which were assumed to be similar across all datasets, a global prior

distribution was generated from the overlap in parameter posterior distributions of all individual fits. Prior distributions for “local” parameters, which were fitted individually to each dataset, were derived from best values of a parameter scan. Only particles contributed to the prior distributions that had the chosen model topology. During SMC ABC, each particle contained values for global parameters, plus one value of the locally changing parameter for each dataset. The distance measures were calculated individually for each dataset and their sum was used to determine the global agreement of simulations with the datasets.

4.8 Statistical analysis

Statistical analysis and curve-fitting was performed in MATLAB and R.

4.8.1 Fitting of distributions

A four-parameter Hill-equation (equation 8) was used to fit dose-response curves using the “nls” function in R. For high-content imaging experiments, the inverse of the standard deviation over all cells was used as weights.

$$f(x) = A \cdot 1/\left(1 + \left(\frac{x}{EC_{50}}\right)^{-n}\right) + c \quad \textbf{(8)}$$

A … Amplitude
n … Hill coefficient
c … Offset

Normal distributions and lognormal distributions were fitted to intensity distributions using the fminsearchbnd[7] and lsqnonlin functions in MATLAB, respectively. A robust linear least-squares fit to relate transcription site intensities at different excitation settings was performed with the fit function in MATLAB.

4.8.2 Histogram matching

Histograms of transcription site intensities of smRNA-FISH and live-cell data were matched by fitting the scale factor between both intensities through minimizing the squared error of the cumulative intensity distributions with the fminsearch function in MATLAB. Only spots with at least 10 RNAs were used for matching of intensity distributions.

4.8.3 Separation of intrinsic and extrinsic noise

Before calculation of noise, the mean of the background signal (quantified from the center of the nucleus from all cells and time points) was subtracted from all data points. Total

[7] John D'Errico, https://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd--fminsearchcon (February 2014)

noise was then calculated as the squared coefficient of variation (CV^2 = variance/mean^2) of all data points across cells and time points.

$$CV_{tot}^2 = \frac{var}{mean^2} = \frac{\frac{1}{NT-1}\sum_{n=1}^{N}\sum_{t=1}^{T}(i_{n,t} - \langle i \rangle)^2}{\langle i \rangle^2} \quad \textbf{(9)}$$

N … Number of cells
T … Number of time points
$i_{n,t}$ … Intensity of cell n at time t
$\langle i \rangle$ … Mean intensity over all cells and time points

Extrinsic noise was approximated as the CV^2 of the area under curve (AUC = sum over all time points for each cell).

$$CV_{ext}^2 = \frac{\frac{1}{N-1}\sum_{n=1}^{N}(AUC_n - \langle AUC \rangle)^2}{\langle AUC \rangle^2} \quad \textbf{(10)}$$

AUC_n … Summed intensity of cell n
$\langle AUC \rangle$ … Mean of summed intensities

Because noise terms are additive (Singh & Soltani 2013), intrinsic noise was then calculated as the difference of total and extrinsic noise.

$$CV_{int}^2 = CV_{tot}^2 - CV_{ext}^2 \quad \textbf{(11)}$$

5 Bibliography

Abremski K, Hoess R. 1984. Bacteriophage p1 site-specific recombination. purification and properties of the cre recombinase protein. *J. Biol. Chem.* 259(3):1509–14

Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, et al. 2017. Rna targeting with crispr-cas13. *Nature*

Annibale P, Gratton E. 2015. Single cell visualization of transcription kinetics variance of highly mobile identical genes using 3d nanoimaging. *Sci. Rep.* 5:9258

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14(3):283–91

Balaban NQ. 2004. Bacterial persistence as a phenotypic switch. *Science (80-.).* 305(5690):1622–25

Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, et al. 2006. Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 38(6):636–43

Bartman CR, Hsu SC, Hsiung CCS, Raj A, Blobel GA. 2016. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol. Cell*. 62(2):237–47

Battich N, Stoeger T, Pelkmans L. 2015. Control of transcript variability in single mammalian cells. *Cell*. 163(7):1596–1610

Baumgärtner S. 2017. *Experimental and theoretical investigation of estrogen dependent transcription in single cells*. University Mainz

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate bayesian computation in population genetics. *Genetics*. 162(4):2025–35

Bergström P, Edlund O. 2014. Robust registration of point sets using iteratively reweighted least squares. *Comput. Optim. Appl.* 58(3):543–61

Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM. 1998. Localization of ash1 mrna particles in living yeast. *Mol. Cell*. 2(4):437–45

Biterge B, Schneider R. 2014. Histone variants: key players of chromatin. *Cell Tissue Res.* 356(3):457–66

Bock C, Beerman I, Lien W-H, Smith ZD, Gu H, et al. 2012. Dna methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol. Cell*. 47(4):633–47

Brock A, Chang H, Huang S. 2009. Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet.* 10(5):336–42

Budnik B, Levy E, Slavov N. 2017. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *bioRxiv*, pp. 1–16

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 523(7561):486–90

Burrill DR, Silver PA. 2010. Making cellular memories. *Cell*. 140(1):13–18

Cai L, Dalal CK, Elowitz MB. 2008. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*. 455(7212):485–90

Capp JP. 2005. Stochastic gene expression, disruption of tissue averaging effects and cancer as a disease of development

Carey LB, van Dijk D, Sloot PMA, Kaandorp JA, Segal E. 2013. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.* 11(4):

Carlberg C. 2010. The impact of transcriptional cycling on gene regulation. *Transcription*. 1(1):13–16

Carr FE, Wong NC. 1994. Characteristics of a negative thyroid hormone response element. *J. Biol. Chem.* 269(6):4175–79

Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. 2008. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*. 453(7194):544–47

Chao J a, Patskovsky Y, Almo SC, Singer RH. 2008. Structural basis for the coevolution of a viral rna-protein complex. *Nat. Struct. Mol. Biol.* 15(1):103–5

Chazaud C, Yamanaka Y, Pawson T, Rossant J. 2006. Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the grb2-mapk pathway. *Dev. Cell*. 10(5):615–24

Chen Y, Sprung R, Tang Y, Ball H, Sangras B, et al. 2007. Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol. Cell. Proteomics*. 6(5):812–19

Chomczynski P. 1993. A reagent for the single-step simultaneous isolation of rna, dna and proteins from cell and tissue samples. *Biotechniques*. 15(3):532–34, 536–37

Chomczynski P, Mackey K, Drews R, Wilfinger W. 1997. Dnazol: a reagent for the rapid isolation of genomic dna. *Biotechniques*. 22(3):550–53

Chong S, Chen C, Ge H, Xie XS. 2014. Mechanism of transcriptional bursting in bacteria. *Cell*. 158(2):314–26

Chow J, Heard E. 2009. X inactivation and the complexities of silencing a sex chromosome. *Curr. Opin. Cell Biol.* 21(3):359–66

Chubb JR, Trcek T, Shenoy SM, Singer RH. 2006. Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16(10):1018–25

Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, et al. 2008. Dynamic proteomics of individual cancer cells in response to a drug. *Sci. (New York, NY)*. 322(5907):1511–16

Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. 2013. Multiplex genome engineering using crispr/cas systems. *Science*. 339(6121):819–23

Corrigan AM, Tunnacliffe E, Cannon D, Chubb JR. 2016. A continuum model of transcriptional bursting. *Elife*. 5:1–38

Dahl C, Grønbæk K, Guldberg P. 2011. Advances in dna methylation: 5-hydroxymethylcytosine revisited

Dalvai M, Bystricky K. 2010. Cell cycle and anti-estrogen effects synergize to regulate cell proliferation and er target gene expression. *PLoS One*. 5(6):

Dar RD, Razooky BS, Singh A, Trimeloni T V., McCollum JM, et al. 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 109(43):17454–59

Darwin C. 1869. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London. 5th ed.

Darzacq X, Yao J, Larson DR, Causse SZ, Bosanac L, et al. 2009. Imaging transcription in living cells. *Annu. Rev. Biophys.* 38:173–96

das Neves RP, Jones NS, Andreu L, Gupta R, Enver T, Iborra FJ. 2010. Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol.* 8(12):e1000560

Deroo BJ, Korach KS. 2006. Estrogen receptors and human disease. *J. Clin. Invest.* 116(3):561–70

Deschênes J, Bourdeau V, White JH, Mader S. 2007. Regulation of greb1 transcription by estrogen receptor alpha through a multipartite enhancer spread over 20 kb of upstream flanking sequences. *J. Biol. Chem.* 282(24):17335–39

Dey SS, Foley JE, Limsirichai P, Schaffer D V., Arkin AP. 2015. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol. Syst. Biol.* 11(5):806–806

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 485(7398):376–80

Doherty CJ, Kay SA. 2010. Circadian control of global gene expression patterns. *Annu. Rev. Genet.* 44(1):419–44

Elowitz MB, Leibler S. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*. 403(6767):335–38

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. *Science (80-.).* 297(5584):1183–86

Femino a M, Fay FS, Fogarty K, Singer RH. 1998. Visualization of single rna transcripts in situ. *Science*. 280(5363):585–90

Filippi S, Barnes CP, Kirk PDW, Kudo T, Kunida K, et al. 2016. Robustness of mek-erk dynamics and origins of cell-to-cell variability in mapk signaling. *Cell Rep.* 15(11):2524–35

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise minimization in eukaryotic gene expression. *PLoS Biol.* 2(6):e137

Fujita K, Iwaki M, Yanagida T. 2016. Transcriptional bursting is intrinsically caused by interplay between rna polymerases on dna. *Nat. Commun.* 7:13788

Fukaya T, Lim B, Levine M. 2016. Enhancer control of transcriptional bursting. *Cell*. 166(2):358–68

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*. 462(7269):58–64

Fusco D, Accornero N, Lavoie B, Shenoy SM, Blanchard JM, et al. 2003. Single mrna molecules demonstrate probabilistic movement in living mammalian cells. *Curr. Biol.* 13(2):161–67

Gandhi SJ, Zenklusen D, Lionnet T, Singer RH. 2011. Transcription of functionally related constitutive genes is not coordinated. *Nat. Struct. Mol. Biol.* 18(1):27–34

Ghosh MG, Thompson DA, Weigel RJ. 2000. Pdzk1 and greb1 are estrogen-regulated genes expressed in hormone-responsive breast cancer. *Cancer Res.* 60(22):6367–75

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81(25):2340–61

Golding I, Paulsson J, Zawilski SM, Cox EC. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*. 123(6):1025–36

Gomez D, Shankman LS, Nguyen AT, Owens GK. 2013. Detection of histone modifications at specific gene loci in single cells in histological sections. *Nat. Methods*

Göttlicher M, Heck S, Herrlich P. 1998. Transcriptional cross-talk, the second mode of steroid hormone receptor action

Green MR, Sambrook J. 2012. *Molecular Cloning: A Laboratory Manual (Fourth Edition)*. Cold Spring Harbor Laboratory Press; 4th edition (June 15, 2012). http://www.amazon.com/Molecular-Cloning-Laboratory-Edition-Three/dp/1936113422

Greer EL, Shi Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* 13(5):343–57

Gretton A, Borgwardt KM, Rasch MJ, Schoelkopf B, Smola A. 2012. A kernel two-sample test. *J. Mach. Learn. Res.* 13:723–73

Groeneweg FL, Van Royen ME, Fenz S, Keizer VIP, Geverts B, et al. 2014. Quantitation of glucocorticoid receptor dna-binding dynamics by single-molecule microscopy and frap. *PLoS One*. 9(3):1–12

Gruber CJ, Tschugguel W, Schneeberger C, Huber JC. 2002. Production and actions of estrogens. *N. Engl. J. Med.* 346(5):340–52

Halpern KB, Caspi I, Lemze D, Elinav E, Ulitsky I, et al. 2015. Nuclear retention of mrna in mammalian tissues report nuclear retention of mrna in mammalian tissues. *CellReports*, pp. 1–10

Hao N, O'Shea EK. 2011. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nat. Struct. Mol. Biol.* 19(1):31–39

Harper C V, Finkenstädt B, Woodcock DJ, Friedrichsen S, Semprini S, et al. 2011. Dynamic analysis of stochastic transcription cycles. *PLoS Biol.* 9(4):e1000607

Helsen C, Kerkhofs S, Clinckemalie L, Spans L, Laurent M, et al. 2012. Structural basis for nuclear hormone receptor dna binding. *Mol. Cell. Endocrinol.* 348(2):411–17

Hocine S, Raymond P, Zenklusen D, Chao JA, Singer RH. 2012. Single-molecule analysis of gene expression using two-color rna labeling in live yeast. *Nat. Methods*. 10(2):119–21

Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, et al. 2012. Noise-mean relationship in mutated promoters. *Genome Res.* 22(12):2409–17

Horvath P, Barrangou R. 2010. Crispr/cas, the immune system of bacteria and archaea. *Science*. 327(5962):167–70

Huh D, Paulsson J. 2011. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* 43(2):95–100

Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. 1997. Molecular reconstruction of sleeping beauty, a tc1-like transposon from fish, and its transposition in human cells. *Cell*. 91(4):501–10

Jaqaman K, Loerke D, Mettlen M, Kuwata H, Grinstein S, et al. 2008. Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods*. 5(8):695–702

JavanMoghadam S, Weihua Z, Hunt KK, Keyomarsi K. 2016. Estrogen receptor alpha is cell cycle-regulated and regulates the cell cycle in a ligand-dependent fashion. *Cell Cycle*. 15(12):1579–90

Jenuwein T, Allis CD. 2001. Translating the histone code. *Science (80-.).* 293(5532):1074–80

Johnston IG, Gaal B, das Neves RP, Enver T, Iborra FJ, Jones NS. 2012. Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Comput. Biol.* 8(3):35–37

Jonkers I, Lis JT. 2015. Getting up to speed with transcription elongation by rna polymerase ii. *Nat Rev Mol Cell Biol*. 16(3):167–77

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6(6):451–64

Kangaspeska S, Stride B, Métivier R, Polycarpou-Schwarz M, Ibberson D, et al. 2008. Transient cyclical methylation of promoter dna. *Nature*. 452(7183):112–15

Karpova TS, Kim MJ, Spriet C, Nalley K, Stasevich TJ, et al. 2008. Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter. *Science*. 319(5862):466–69

Kebede AF, Schneider R, Daujat S. 2015. Novel types and sites of histone modifications emerge as players in the transcriptional regulation contest. *FEBS J.* 282(9):1658–74

Kempe H, Schwabe A, Crémazy F, Verschure PJ, Bruggeman FJ. 2015. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Mol. Biol. Cell*. 26(4):797–804

Kim J, Marioni JC. 2013. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome Biol.* 14(1):R7

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 161(5):1187–1201

Ko MSH, Nakauchi H, Takahashi N. 1990. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *EMBO J.* 9(9):2835–42

Kouzarides T. 2007. Chromatin modifications and their function. *Cell*. 128(4):693–705

Kreso A, O'Brien CA, van Galen P, Gan OI, Notta F, et al. 2013. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science (80-.).* 339(6119):543–48

Lane K, Van Valen D, DeFelice MM, Macklin DN, Kudo T, et al. 2017. Measuring signaling and rna-seq in the same cell links gene expression to dynamic patterns of nf-κb activation. *Cell Syst.* 4(4):458–69.e5

Larson D, Zenklusen D, Wu B, Chao J, Singer R. 2011. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science (80-.).*

Larson DR, Fritzsch C, Sun L, Meng X, Lawrence DS, Singer RH. 2013. Direct observation of frequency modulated transcription in single cells using light activation. *Elife*. 2:e00750–e00750

Lee TI, Young R a. 2000. Transcription of eukaryotic protein- coding genes. *Annu. Rev. Genet.*, pp. 77–137

Lee A V., Oesterreich S, Davidson NE. 2015. Mcf-7 cells - changing the course of breast cancer research and care for 45 years

Lemaire V, Lee C, Lei J, Métivier R, Glass L. 2006. Sequential recruitment and combinatorial assembling of multiprotein complexes in transcriptional activation. *Phys. Rev. Lett.* 96(19):2–5

Lenstra TL, Rodriguez J, Chen H, Larson DR. 2016. Transcription dynamics in living cells. *Annu. Rev. Biophys.* 45(1):25–47

Levesque MJ, Raj A. 2013. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods*. 10(3):246–48

Li E, Beard C, Jaenisch R. 1993. Role for dna methylation in genomic imprinting. *Nature*. 366(6453):362–65

Lin Y, Sohn CH, Dalal CK, Cai L, Elowitz MB. 2015. Combinatorial gene regulation by modulation of relative pulse timing. *Nature*. 527(7576):54–58

Lipovka Y, Konhilas JP. 2016. The complex nature of oestrogen signalling in breast cancer: enemy or ally? *Biosci. Rep.* 36(3):e00352–e00352

Lombard-Banek C, Moody SA, Nemes P. 2016. Single-cell mass spectrometry for discovery proteomics: quantifying translational cell heterogeneity in the 16-cell frog (xenopus) embryo. *Angew. Chemie Int. Ed.* 55(7):2454–58

Lubeck E, Cai L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods*. 9(7):743–48

Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. 2014. Single-cell in situ rna profiling by sequential hybridization. *Nat. Methods*. 11(4):360–61

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*. 389(6648):251–60

Lumachi F, Brunello A, Maruzzo M, Basso U, Basso SMM. 2013. Treatment of estrogen receptor-positive breast cancer. *Curr. Med. Chem.* 20(5):596–604

Lungu C, Pinter S, Broche J, Rathert P, Jeltsch A. 2017. Modular fluorescence complementation sensors for live cell detection of epigenetic signals at endogenous genomic sites. *Nat. Commun.* 8(1):649

Mao C, Brown CR, Falkovskaia E, Dong S, Hrabeta-Robinson E, et al. 2010. Quantitative analysis of the transcription control mechanism. *Mol. Syst. Biol.* 6:

Martelotto LG, Ng CK, Piscuoglio S, Weigelt B, Reis-Filho JS. 2014. Breast cancer intra-tumor heterogeneity. *Breast Cancer Res.* 16(3):210

Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, et al. 2009. Molecular evolution of a novel hyperactive sleeping beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* 41(6):753–61

McDonnell DP, Norris JD. 2002. Connections and regulation of the human estrogen receptor. *Science*. 296(5573):1642–44

McGranahan N, Swanton C. 2017. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*. 168(4):613–28

McNally JG. 2000. The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science (80-.).* 287(5456):1262–65

Métivier R, Gallais R, Tiffoche C, Le Péron C, Jurkowska RZ, et al. 2008. Cyclical dna methylation of a transcriptionally active promoter. *Nature*. 452(7183):45–50

Métivier R, Penot G, Hübner MR, Reid G, Brand H, et al. 2003. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*. 115(6):751–63

Métivier R, Reid G, Gannon F. 2006. Transcription in four dimensions: nuclear receptor-directed initiation of gene expression. *EMBO Rep.* 7(2):161–67

Mohammed H, D'Santos C, Serandour A a., Ali HR, Brown GD, et al. 2013. Endogenous purification reveals greb1 as a key estrogen receptor regulatory factor. *Cell Rep.* 3(2):342–49

Molina N, Suter DM, Cannavo R, Zoller B, Gotic I, Naef F. 2013. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Natl. Acad. Sci. U. S. A.* 110(51):20563–68

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, et al. 2013. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*. 502(7469):59–64

Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, et al. 2016. Programmable rna tracking in live cells with crispr/cas9. *Cell*. 165(2):488–96

Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013. Systematic identification of signal-activated stochastic gene regulation. *Science (80-.).* 339(6119):584–87

Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. 2006. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*. 441(7095):840–46

Nicolas D, Phillips NE, Naef F. 2017. What shapes eukaryotic transcriptional bursting? *Mol. BioSyst.* 13(7):1280–90

Nowak SJ, Corces VG. 2004. Phosphorylation of histone h3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet.* 20(4):214–20

O'Lone R, Frith MC, Karlsson EK, Hansen U. 2004. Genomic targets of nuclear estrogen receptors. *Mol. Endocrinol.* 18(8):1859–75

Ochiai H, Sugawara T, Sakuma T, Yamamoto T. 2014. Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells. *Sci. Rep.* 4:7125

Ochiai H, Sugawara T, Yamamoto T. 2015. Simultaneous live imaging of the transcription and nuclear position of specific genes. *Nucleic Acids Res.* 43(19):1–12

Otsu N. 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man. Cybern.* 9(1):62–66

Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, et al. 2015. Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. *Mol. Cell*. 58(2):339–52

Paek AL, Liu JC, Loewer A, Forrester WC, Lahav G. 2016. Cell-to-cell variation in p53 dynamics leads to fractional killing. *Cell*. 165(3):631–42

Parada LA, Sotiriou S, Misteli T. 2004. Spatial genome organization. *Exp. Cell Res.* 296(1):64–70

Park K-J, Krishnan V, O'Malley BW, Yamamoto Y, Gaynor RB. 2005. Formation of an ikkα-dependent transcription complex is required for estrogen receptor-mediated gene activation. *Mol. Cell*. 18(1):71–82

Patil VS, Zhou R, Rana TM. 2014. Gene regulation by non-coding rnas. *Crit. Rev. Biochem. Mol. Biol.* 49(1):16–32

Paulsson J. 2004. Summing up the noise in gene networks. *Nature*. 427(6973):415–18

Paulsson J. 2005. Models of stochastic gene expression. *Phys. Life Rev.* 2(2):157–75

Peccoud J, Ycart B. 1995. Markovian modeling of gene-product synthesis

Perkins TJ, Swain PS. 2009. Strategies for cellular decision-making. *Mol. Syst. Biol.* 5:

Picaud S, Da Costa D, Thanasopoulou A, Filippakopoulos P, Fish P V., et al. 2013. Pfi-1, a highly selective protein interaction inhibitor, targeting bet bromodomains. *Cancer Res.* 73(11):3336–46

Poorey K, Viswanathan R, Carver MN, Karpova TS, Cirimotich SM, et al. 2013. Measuring chromatin interaction dynamics on the second time scale at single-copy genes. *Science*. 342(6156):369–72

Prall OW, Rogan EM, Sutherland RL. 1998. Estrogen regulation of cell cycle progression in breast cancer cells. *J. Steroid Biochem. Mol. Biol.* 65(1-6):169–74

Rae JM, Johnson MD, Scheys JO, Cordero KE, Larios JM, Lippman ME. 2005. Greb1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Res. Treat.* 92:141–49

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. 2006. Stochastic mrna synthesis in mammalian cells. *PLoS Biol.* 4(10):e309

Raj A, van den Bogaard P, Rifkin S a, van Oudenaarden A, Tyagi S. 2008. Imaging individual mrna molecules using multiple singly labeled probes. *Nat. Methods*. 5(10):877–79

Ran F, Hsu P, Wright J, Agarwala V. 2013. Genome engineering using the crispr-cas9 system. *Nat. Protoc.* 8(11):2281–2308

Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. *Science*. 304(5678):1811–14

Reid G, Hübner MR, Métivier R, Brand H, Denger S, et al. 2003. Cyclic, proteasome-mediated turnover of unliganded and liganded eralpha on responsive promoters is an integral feature of estrogen signaling. *Mol. Cell.* 11(3):695–707

Reid G, Metivier R, Lin CY, Denger S, Ibberson D, et al. 2005. Multiple mechanisms induce transcriptional silencing of a subset of genes, including oestrogen receptor alpha, in response to deacetylase inhibition by valproic acid and trichostatin a. *Oncogene*. 24(31):4894–4907

Rigaud G, Roux J, Pictet R, Grange T. 1991. In vivo footprinting of rat tat gene: dynamic interplay between the glucocorticoid receptor and a liver-specific factor. *Cell*. 67(5):977–86

Robinson PJ, Rhodes D. 2006. Structure of the "30nm" chromatin fibre: a key role for the linker histone. *Curr. Opin. Struct. Biol.* 16(3):336–43

Robinson-Rechavi M. 2003. The nuclear receptor superfamily. *J. Cell Sci.* 116(4):585–86

Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. 1998. Dna double-stranded breaks induce histone h2ax phosphorylation on serine 139. *J. Biol. Chem.* 273(10):5858–68

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, et al. 2015. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*

Roux J, Hafner M, Bandara S, Sims JJ, Hudson H, et al. 2015. Fractional killing arises from cell-to-cell variability in overcoming a caspase activity threshold. *Mol. Syst. Biol.* 11(5):803–803

Rybakova KN, Bruggeman FJ, Tomaszewska A, Moné MJ, Carlberg C, Westerhoff H V. 2015. Multiplex eukaryotic transcription (in)activation: timing, bursting and cycling of a ratchet clock mechanism. *PLOS Comput. Biol.* 11(4):e1004236

Sanchez A, Choubey S, Kondev J. 2013. Stochastic models of transcription: from single molecules to single cells. *Methods*

Schmiedel JM, Klemm SL, Zheng Y, Sahay A, Bluthgen N, et al. 2015. Microrna control of protein expression noise. *Science (80-.).* 348(6230):128–32

Scholes C, DePace AH, Sánchez Á. 2017. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Syst.*

Schwabe A, Rybakova KN, Bruggeman FJ. 2012. Transcription stochasticity of complex gene regulation models. *Biophys. J.* 103(6):1152–61

Senecal A, Munsky B, Proux F, Ly N, Braye FE, et al. 2014. Transcription factors modulate c-fos transcriptional bursts. *Cell Rep.* 8(1):75–83

Shang Y, Hu X, DiRenzo J, Lazar M a, Brown M. 2000. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell*. 103(6):843–52

Sharma S V., Lee DY, Li B, Quinlan MP, Takahashi F, et al. 2010. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*. 141(1):69–80

Sharova L V., Sharov AA, Nedorezov T, Piao Y, Shaik N, Ko MSH. 2009. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.* 16(1):45–58

Sherman MS, Lorenz K, Lanier MH, Cohen BA. 2015. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell Syst.* 1(5):315–25

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15(4):272–86

Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, et al. 2006. Variability and memory of protein levels in human cells. *Nature*. 444(7119):643–46

Singh A, Razooky B, Cox CD, Simpson ML, Weinberger LS. 2010. Transcriptional bursting from the hiv-1 promoter is a significant source of stochastic noise in hiv-1 gene expression. *Biophys. J.* 98(8):L32–34

Singh A, Soltani M. 2013. Quantifying intrinsic and extrinsic variability in stochastic gene expression models. *PLoS One*. 8(12):

Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, Golding I. 2016. Single-cell analysis of transcription kinetics across the cell cycle. *Elife*. 5(JANUARY2016):1–24

Skupsky R, Burnett JC, Foley JE, Schaffer D V., Arkin AP. 2010. Hiv promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS Comput. Biol.* 6(9):

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, et al. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*. 11(8):817–20

Sonnen KF, Aulehla A. 2014. Dynamic signal encoding-from cells to organisms

Soule HD, Vazguez J, Long A, Albert S, Brennan M. 1973. A human cell line from a pleural effusion derived from a breast carcinoma. *J. Natl. Cancer Inst.* 51(5):1409–16

Stewart-Ornstein J, Weissman JS, El-Samad H. 2012. Cellular noise regulons underlie fluctuations in saccharomyces cerevisiae. *Mol. Cell*. 45(4):483–93

Stoeger T, Battich N, Herrmann MD, Yakimovich Y, Pelkmans L. 2015. Computer vision for image-based transcriptomics. *Methods*. 85:44–53

Stoeger T, Battich N, Pelkmans L. 2016. Passive noise filtering by cellular compartmentalization. *Cell*. 164(6):1151–61

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature*. 403(6765):41–45

Sun J, Nawaz Z, Slingerland JM. 2007. Long-range activation of greb1 by estrogen receptor via three distal consensus estrogen-responsive elements in breast cancer cells. *Mol. Endocrinol.* 21(11):2651–62

Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. 2011. Mammalian genes are transcribed with widely different bursting kinetics. *Science*. 332(6028):472–74

Swain PS, Elowitz MB, Siggia ED. 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 99(20):12795–800

Tan M, Luo H, Lee S, Jin F, Yang J, et al. 2011. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*. 146(6):1016–28

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. Mrna-seq whole-transcriptome analysis of a single cell. *Nat. Methods*. 6(5):377–82

Tantale K, Mueller F, Kozulic-Pirher A, Lesne A, Victor J-M, et al. 2016. A single-molecule view of transcription reveals convoys of rna polymerases and multi-scale bursting. *Nat. Commun.* 7:12248

Thattai M, van Oudenaarden A. 2001. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci.* 98(15):8614–19

Toni T, Stumpf MPH. 2009. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*. 26(1):104–10

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*. 6(31):187–202

Vantaggiato C, Tocchetti M, Cappelletti V, Gurtner A, Villa A, et al. 2014. Cell cycle dependent oscillatory expression of estrogen receptor-α links pol ii elongation to neoplastic transformation. *Proc. Natl. Acad. Sci.* 111(26):9561–66

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10(4):252–63

Vinuelas J, Kaneko G, Coulon A, Vallin E, Morin V, et al. 2013. Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC Biol.* 11(1):15

Viñuelas J, Kaneko G, Coulon A, Beslon G, Gandrillon O. 2012. Towards experimental manipulation of stochasticity in gene expression. *Prog. Biophys. Mol. Biol.* 110(1):44–53

wa Maina C, Honkela A, Matarese F, Grote K, Stunnenberg HG, et al. 2014. Inference of rna polymerase ii transcription dynamics from chromatin immunoprecipitation time course data. *PLoS Comput. Biol.* 10(5):1–17

Wakeling AE, Dukes M, Bowler J. 1991. A potent specific pure antiestrogen with clinical potential. *Cancer Res.* 51(15):3867–73

Walsh C, Chaillet J, Bestor T. 1998. Transcription of iap endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* 20(October):116–17

Walters MC, Fiering S, Eidemiller J, Magis W, Groudine M, Martin DI. 1995. Enhancers increase the probability but not the level of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 92(15):7125–29

Wang C. 2001. Direct acetylation of the estrogen receptor alpha hinge region by p300 regulates transactivation and hormone sensitivity. *J. Biol. Chem.* 276(21):18375–83

Wang GG, Allis CD, Chi P. 2007. Chromatin remodeling and cancer, part ii: atp-dependent chromatin remodeling. *Trends Mol. Med.* 13(9):373–80

Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci.* 108(16):E67–76

Ward HW. 1973. Anti-oestrogen therapy for breast cancer: a trial of tamoxifen at two dose levels. *Br. Med. J.* 1(5844):13–14

Watters JJ, Campbell JS, Cunningham MJ, Krebs EG, Dorsa DM. 1997. Rapid membrane effects of steroids in neuroblastoma cells: effects of estrogen on mitogen activated protein kinase signalling cascade and c-fos immediate early gene transcription. *Endocrinology*. 138(9):4030–33

Weigel NL. 1996. Steroid hormone receptors and their regulation by phosphorylation. *Biochem. J.* 319:657–67

Weintraub H. 1988. Formation of stable transcription complexes as assayed by analysis of individual templates. *Proc. Natl. Acad. Sci. U. S. A.* 85(16):5819–23

Wu B, Chao J a, Singer RH. 2012. Fluorescence fluctuation spectroscopy enables quantitative imaging of single mrnas in living cells. *Biophys. J.* 102(12):2936–44

Yamaga R, Ikeda K, Horie-Inoue K, Ouchi Y, Suzuki Y, Inoue S. 2013. Rna sequencing of mcf-7 breast cancer cells identifies novel estrogen-responsive genes with functional estrogen receptor-binding sites in the vicinity of their transcription start sites. *Horm. Cancer*. 4(4):222–32

Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. 2013. One-step generation of mice carrying reporter and conditional alleles by crispr/cas-mediated genome engineering. *Cell*. 154(6):1370–79

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-blast: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 13:134

Yun M, Wu J, Workman JL, Li B. 2011. Readers of histone modifications. *Cell Res.* 21(4):564–78

Zenklusen D, Larson DR, Singer RH. 2008. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* 15(12):1263–71

Zhang J, Chen L, Zhou T. 2012. Analytical distribution and tunability of noise in a model of promoter progress. *Biophys. J.* 102(6):1247–57

Zoller B, Nicolas D, Molina N, Naef F. 2015. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol. Syst. Biol.* 11(7):823–823

Zopf CJ, Quinn K, Zeidman J, Maheshri N. 2013. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput. Biol.* 9(7):1–12

Appendix

List of tables

List of figures

List of abbreviations

ActD	Actinomycin D
ATP	Adenosine triphosphate
BFP2	Blue fluorescent protein 2
bp	Base pair
CMV	Cytomegalovirus
CRISPR	Clustered regularly interspaced short palindromic repeat
Ct	Threshold cycle
DAPI	4′,6-Diamidin-2-phenylindol
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
E_2	17β-estradiol
EC_{50}	Half maximal effective concentration
ERα	Estrogen receptor α
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FISH	Fluorescence *in-situ* hybridization
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GFP	Green fluorescent protein
GLOX	Glucose oxidase
GREB1	Growth regulation by estrogen in breast cancer 1
HAT	Histone acetyl transferase
HDAC	Histone deacetylase
HDR	Homology directed repair
HMT	Histone methyl transferase
IC_{50}	Half maximal inhibitory concentration
ICI	ICI 182,780 (Fulvestrant)
IRES	Internal ribosomal entry site
kb	Kilobase
loxP	Locus of crossing over in bacteriophage P1
MS2	Bacteriophage MS2
NaBu	Sodium butyrate
NAD+	Nicotinamide adenine dinucleotide
ORF	Open reading frame
OHT	4-Hydroxy-Tamoxifen
PBS	Phosphate-buffered saline
PCP	PP7 coat protein
PCR	Polymerase chain reaction
PFA	Para-formaldehyde
PP7	*Pseudomonas* phage 7
px	Pixel
qPCR	Quantitative PCR
RNA	Ribonucleic acid
RT	Reverse transcriptase
sCMOS	Scientific complementary metal–oxide–semiconductor
SERM	Selective estrogen receptor modulator
SWI/SNF	Switch/Sucrose Non-Fermentable
tdGFP	Tandem-dimer of GFP
tdPCP	Tandem-dimer of PCP
TSA	Trichostatin A

Acknowledgements

Curriculum Vitae