From the Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

PRINCIPLES OF TRANSCRIPTIONAL BURSTING IN MAMMALIAN CELLS

Anton Larsson

Karolinska Institutet

Stockholm 2023

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91-8017-035-2

Principles of transcriptional bursting in mammalian cells

Thesis for Doctoral Degree (Ph.D.)

By

Anton Larsson

The thesis will be defended in public at Eva & Georg Klein, Biomedicum, Solnavägen 9, Solna, Friday 16 June 2023, 13:00

Principal Supervisor:
Professor Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology

Co-supervisor(s):
Professor Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Yishao Zhou
Stockholm University
Department of Mathematics

Opponent:
Professor Alexander van Oudenaarden
Hubrecht Institute

Examination Board:
Assistant Professor Martin Enge
Karolinska Institutet
Department of Oncology-Pathology

Professor Tuuli Lappalainen
Kungliga Tekniska Högskolan
School of Engineering Sciences in Chemistry, Biotechnology and Health
Division of Gene Technology

Professor Jonas Muhr
Karolinska Institutet
Department of Cell and Molecular Biology

Till min pappa

Popular science summary of the thesis

Transcription is a fundamental process that occurs in all cells of the human body and allows us to use the genes we inherit. It "transcribes" the DNA (genes) in our cell's nucleus into RNA which can further be used to make protein. Depending on which genes are transcribed, the cell can perform different tasks with the proteins they make. The ability to control which genes are being transcribed is crucial to the development of the individual and largely responsible for the incredible diversity of cells in the human body.

Transcription occurs in discrete bursts, with many RNA molecules being produced in a short period of time followed by a period of inactivity. Studying how transcriptional bursting is controlled and its consequences is important. However, most methods that have been used to measure transcription hide this phenomenon and the ones that can be used to observe transcriptional bursting can only be applied to one or a few genes at a time. Furthermore, we have two copies of our chromosomes (molecules of DNA), and transcription occurs independently from each gene copy.

In this thesis, I describe my effort to develop new methods to measure transcriptional bursts and findings related to them. All the studies have used a relatively new and quickly developing method called single-cell RNA sequencing. This method turns the cell's RNA into DNA, which is then sequenced to determine which genes have been transcribed by that cell.

In **Paper I,** I developed a computational model that allowed us to estimate the transcriptional bursting parameters of each transcribed gene. These parameters can be summarized into two characteristics, the frequency of bursts (burst frequency) and the average number of transcripts produced in one burst (burst size). This model was then applied to single-cell RNA sequencing data from a mouse. By using single-cell RNA sequencing data we can study the transcriptional bursting behavior of many genes simultaneously. We showed that certain parts of the DNA which regulate transcription affect transcriptional bursting in different ways. Parts of the DNA which are called enhancers can help increase burst frequency, while other parts known as promoters instead affect burst size. The mice we used is a crossbreed between two distantly related mice which allows us to distinguish between the two copies of each chromosome since they have lots of differences in their DNA. In **Paper II** we show that the frequency of transcriptional bursts determines how often we observe either copy of a gene. This finding can help explain how some genetic diseases have variable penetrance on patients.

The biological sex of an individual is determined by the presence of the Y chromosome. Males have one Y chromosome and one X chromosome, while females have two X chromosomes. In any other case having only one copy of a chromosome is not

compatible with life. It turns out that the X chromosome is peculiar in multiple ways. Interestingly, in females one of their X chromosomes become a condensed molecule early in development and cannot be transcribed from, which is called X chromosome inactivation. In the 1960's the researcher Susumu Ohno hypothesized that transcription from the single X chromosome is boosted to be equal to having two chromosomes. The single active X chromosome is working overtime. In **Paper III** we argue that this is indeed the case, and that is achieved by an increase in the frequency of transcriptional bursts. Furthermore, by studying female cells that are going through X chromosome inactivation, we find that the increase in bursting frequency is reliant on the number of active X chromosomes.

In **Paper IV and V** we describe two methods that allow us to measure what kind of RNA is produced within a specified time-window. This is done by giving cells growing in the culture dish in the lab a building block of RNA that is slightly different from the normal version, but similar enough that the cell uses it during transcription. Using a chemical conversion step and computational algorithms, we can distinguish between the RNA we sequenced that is newly transcribed and the RNA that was transcribed before we added our unusual building block. We can use this to study transcriptional responses, which we demonstrate in Paper IV by stimulating immune cells. Furthermore, in Paper V we make the method better, and show the advantages of studying transcriptional bursting with molecules that are only recently produced.

In conclusion, studying transcriptional bursting is relevant to many topics in cell biology and the studies in this thesis have demonstrated the possibility to study it for many genes at once. This approach can be used to study cells at a deeper level.

Populärvetenskaplig sammanfattning

Transkription är en fundamental process som tar plats i alla celler i människans kropp och låter oss använda generna vi har ärvt. Den ”transkriberar” DNAt (gener) i vår cellkärna till RNA som sedan kan användas för att göra protein. Beroende på vilka gener som transkriberas kan cellen utöva de olika uppgifter som proteinerna kan göra. Förmågan att kontrollera vilka gener som transkriberas är avgörande för utvecklingen av individen och till stor del ansvarig för den enorma mångfalden av olika celler i vår kropp.

Transkription sker i avskilda “explosioner” av aktivitet, som jag med viss motvilja för anglicismer kommer kallar för bursts. Dessa bursts producerar många RNA molekyler som följs av längre perioder av inaktivitet. Att studera hur bursts kontrolleras och dess konsekvenser är viktigt. Men de flesta metoder som används för att mäta transkription döljer detta fenomen, och de metoder som man kan använda kan bara tillämpas på en eller ett fåtal gener åt gången. Dessutom har vi två kopior av våra kromosomer (stora molekyler av DNA) och transkription från dessa sker oberoende från varje genkopia.

Denna avhandling beskriver jag mina insatser att utveckla nya metoder för att mäta bursts och mina upptäcker relaterat till dem. Alla studier har använt en relativt ny och snabbt växande metod som kan RNA-sekvensering av enskilda celler. Denna metod gör om cellens RNA till DNA som sedan kan sekvenseras för att fastställa vilka gener som har transkriberats av den cellen.

I **Delarbete I** beskriver jag en beräkningsmodell för att uppskatta burstparametrar för varje transkriberad gen. Dessa parametrar kan sammanfattas med två kännetäcken, hur ofta en gen burstar (burstfrekvens) och hur många RNA molekyler produceras i en burst i genomsnitt (burststorlek). Denna modell användes sedan på RNA-sekvenseringsdata på enskilda musceller. Genom att använda denna metod kan vi studera bursts från många gener samtidigt. Vi visade att olika delar av vårt DNA som kontrollerar transkription har olika påverkan på bursts. Vi visade att enhancers påverkar burstfrekvensen, medan en annan grupp som kallas promoters påverkar burststorleken. Musen vi använde var en korsning mellan två avlägset besläktade möss vilket låter oss skilja på de två olika genkopiorna på grund av den genetiska variationen i DNAt. **Delarbete II** visar vi att burstfrekvensen bestämmer hur ofta vi ser de två olika genkopiorna. Denna iakttagelse kan förklara varför vissa genetiska sjukdomar har olika penetrans i olika patienter.

Det biologiska könet bestäms av närvaron av Y kromosomen. Män har en Y kromosom och en X kromosom medan kvinnor har två X kromosomer. I andra alla fall så är det inte möjligt att bara ha en kopia av en kromosom. Det visar sig att X kromosomen är speciell på flera sätt. Intressant nog så kondenseras en av kvinnans X kromosomer och görs otillgänglig i alla celler tidigt i utvecklingen, detta kallas X kromosom inaktivering. På

1960-talet formulerade forskaren Susumu Ohno hypotesen att transkriptionen från X kromosomen dessutom sker i dubbel mängd för att matcha samma nivå från två kromosomkopior. Denna ensamma X kromosom jobbade övertid. **Delarbete III** visar för att detta är fallet och uppnås genom att öka burstfrekvensen för X kromosomens gener. Dessutom visar vi genom att studera kvinnliga celler vars X kromosom inaktiveras att denna ökning i burstfrekvens är beroende på antalet aktiva X kromosomer.

I **Delarbete IV och V** beskriver vi två metoder som låter oss mäta vilka gener som transkriberas under ett tidsfönster i enskilda celler. Detta gör vi genom att ge celler som växer i en cellodlingsplatta en byggsten av RNA som är lite annorlunda, men tillräckligt lika att cellen använder den vid transkription. Genom en kemisk omvandling och beräkningsalgoritmer kan vi skilja mellan RNA som transkriberades nyligen och RNA som transkriberades innan vi gav cellerna den annorlunda byggstenen. Vi kan använda denna metod för at studera transkiptionella svar, som vi visar i Delarbete IV genom att stimulera immunceller. Dessutom gör vi metoden bättre i Delarbete V och visar de fördelar denna metod har för att studera bursts genom att avgränsa analysen till det RNA som är nytt.

För att sammanfatta så är det relevant att mäta transkriptionella bursts för många frågor i cellbiologi och delarbeten i denna avhandling visar möjligheten att undersöka flera gener samtidigt i enskilda celler. Detta tillvägagångssätt kan användas för att studera våra celler på en djupare nivå än tidigare.

Abstract

In mammalian cells, transcription occurs in discrete bursts leading to fluctuations in transcripts from expressed genes. Although this behavior was first reported not long after the discovery of messenger RNA (mRNA), the methods to measure transcriptional bursting have been limited in throughput and scalability. To enable transcriptome wide analysis of transcriptional bursting, I have developed multiple methods to estimate transcriptional bursting behavior using deeply sequenced single-cell RNA-sequencing data. In **Paper I**, we use a computational likelihood method based on the two-state model of transcriptional bursting to estimate allele-resolved bursting kinetics of mouse cells. The transcriptome wide estimates allow us to detect how the genomic regions of enhancers and promoters affect transcriptional bursts. To a first approximation, enhancers direct the frequency of bursts while promoters influence the number of transcripts per burst. The fluctuations of the transcript alleles may cause phenotypic variability over time. In **Paper II**, we directly show that the bursting behavior of a gene determines how often monoallelic expression is observed from that gene. Moreover, we show that this can lead to false positive monoallelic observations in bulk experiments if not considered. This can be concluded for the genes present on autosomal chromosomes. The X chromosome, however, has only one active copy in mature cells which causes complications in gene dosage. In **Paper III,** we report that the genes on the single active X chromosome are upregulated compared to the genes on the autosomal chromosomes, and that this upregulation is achieved through an increased burst frequency. Furthermore, this upregulation is coupled to X chromosome inactivation in females. To study transcriptional bursting at a more resolved time scale, we developed novel single cell sequencing methods using metabolic labeling in **Paper IV and V.** These methods supply the nucleotide analog 4-thiouridine to cells during cell culture, which become incorporated during transcription. Due to the alkylation reaction during library preparation leading to the incorporation of the wrong nucleotide during reverse transcription, the incorporated 4-thiouridine can be computationally detected as mismatches to the reference genome during analysis. We use this approach to study responses to a perturbation (Paper IV) and to study transcriptional bursting during a 2-hour time window (Paper V). This data allows the further dissection of transcriptional bursting and the ability to study co-bursting in single cells. We show that the synthesis rate mainly determines burst size and not the transcriptional off rate. We do not find co-bursting to be a general phenomenon across the transcriptome but do find certain gene pairs that exhibit co-bursting.

List of scientific papers

I. **Larsson, Anton J. M.**, Per Johnsson, Michael Hagemann-Jensen*, Leonard Hartmanis*, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. 2019.
"Genomic Encoding of Transcriptional Burst Kinetics." ***Nature*** 565 (7738): 251–54.

II. **Larsson, Anton J. M.**, Christoph Ziegenhain*, Michael Hagemann-Jensen*, Björn Reinius, Tina Jacob, Tim Dalessandri, Gert-Jan Hendriks, Maria Kasper, and Rickard Sandberg. 2021.
"Transcriptional Bursts Explain Autosomal Random Monoallelic Expression and Affect Allelic Imbalance." ***PLOS Computational Biology*** 17 (3): e1008772.

III. **Larsson, Anton J. M.**, Christos Coucoravas, Rickard Sandberg, and Björn Reinius. 2019.
"X-Chromosome Upregulation Is Driven by Increased Burst Frequency." ***Nature Structural & Molecular Biology*** 26 (10): 963–69.

IV. Hendriks, Gert-Jan, Lisa A. Jung, **Anton J. M. Larsson**, Michael Lidschreiber, Oscar Andersson Forsman, Katja Lidschreiber, Patrick Cramer, and Rickard Sandberg. 2019.
"NASC-Seq Monitors RNA Synthesis in Single Cells." ***Nature Communications*** 10 (1): 3138.

V. Hendriks, Gert-Jan*, Daniel Ramsköld*, **Anton J.M. Larsson***, Juliane V. Mayr, Christoph Ziegenhain, Michael Hagemann-Jensen, Leonard Hartmanis, Rickard Sandberg.
Single-cell new RNA sequencing reveals principles of transcription at the resolution of individual bursts. ***Manuscript***.

*Equal contribution.

Scientific papers not included in the thesis

Larsson, Anton J. M., Geoff Stanley, Rahul Sinha, Irving L. Weissman, and Rickard Sandberg. 2018.
"Computational Correction of Index Switching in Multiplexed Sequencing Libraries." ***Nature Methods*** 15 (5): 305–7.

Lee, Woojoo, Arvid Sjölander, **Anton Larsson**, and Yudi Pawitan. 2018.
"Likelihood-Based Inference for Bounds of Causal Parameters." ***Statistics in Medicine*** 37 (30): 4695–4706.

Hagemann-Jensen, Michael, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, **Anton J. M. Larsson**, Omid R. Faridani, and Rickard Sandberg. 2020.
"Single-Cell RNA Counting at Allele and Isoform Resolution Using Smart-Seq3." ***Nature Biotechnology*** 38 (6): 708–14.

Mold, Jeff E., Laurent Modolo, Joanna Hård, Margherita Zamboni, **Anton J. M. Larsson**, Moa Stenudd, Carl-Johan Eriksson, et al. 2021.
"Divergent Clonal Differentiation Trajectories Establish CD8+ Memory T Cell Heterogeneity during Acute Viral Infections in Humans." ***Cell Reports*** 35 (8): 109174.

Contents

List of abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger Ribonucleic acid
smFISH	Single molecule fluorescence in situ hybridization
F1	First generation (with regards to cross-mating)
RT	Reverse transcriptase

1 Introduction

This thesis is about transcription. Specifically, how transcription occurs, transcriptional bursting, my efforts to measure transcriptional bursts and some underlying principles of how transcriptional bursting is controlled. Mammals are very complex organisms, with many organs that work together to constitute the animal. Those organs are in turn composed of cells that are specialized to perform specific tasks. Transcription is important because the main way these cells are able to specialize is dependent on which genes are transcribed.

2 Literature review

2.1 The genome

The genome contains the information needed for the organism to develop and function. It is composed of deoxyribonucleic acid (DNA), that is organized into chromosomes in the nucleus of the cell and circular DNA in the mitochondria. The human diploid genome consists of 23 pairs of chromosomes, of which 22 pairs are autosomal and 1 pair is the sex chromosomes, and was completely sequenced just last year (Nurk et al. 2022).

2.1.1 The genome is packed and structured

The genome has multiple structural features which are important for understanding transcription. The fundamental structural unit of the chromosome is the nucleosome, which consists of a segment of DNA wrapped around a protein complex called a histone. The histone itself is an octamer composed of two copies of four proteins: histone proteins H2A, H2B, H3, and H4 (Babu and Verma 1987). The DNA is wound 1.65 turns around the histone octamer, corresponding to 146 base pairs of DNA (Luger et al. 1997). The complex of nucleosomes is called chromatin. The open form of chromatin is called euchromatin and is associated with the active transcription of genes. Approximately 92% of the human genome is euchromatic (International Human Genome Sequencing Consortium 2004). The closed form of chromatin is called heterochromatin. The spacing between nucleosomes in heterochromatin is much narrower compared to euchromatin and chromatin in the centromeres and near the telomeres are invariably heterochromatic (Saksouk, Simboeck, and Déjardin 2015). Regardless of the cell type, these regions are tightly packed, and no polymerase can access these regions (Volpe et al. 2002). Furthermore, the conformation of chromatin is regulated by histone modifications. Specific histone residues may be chemically modified to affect the structure of the nucleosome, which may promote or prevent transcription. For example,

heterochromatic DNA has been associated with the methylation of H3K9[1] (i.e., H3K9me2 or H3K9me3) (Rosenfeld et al. 2009). Although most histone modifications clearly alter the structure of the chromatin, more research is needed to achieve a full understanding of their effects. Zooming in even further, the chromosomes are also organized into cell-type invariant topologically associated domains. The DNA within a topologically associated domain typically only form physical contacts within their domain (Dixon, Gorkin, and Ren 2016).

2.2 Transcription

While the genome contains the information for virtually all possible tasks, the DNA must be transcribed into ribonucleic acid (RNA) to enable the use of that information.

2.2.1 Polymerase transcribes DNA into RNA

Higher eukaryotes have three polymerases that use DNA as a template to produce RNA. They are aptly named RNA polymerase I, II and III. These RNA polymerases are all multiprotein complexes with 12-17 subunits and transcribe different kinds of RNA. RNA polymerase I only transcribes ribosomal RNA, which is needed for the translation of RNA into protein (Russell and Zomerdijk 2006). RNA polymerase III mostly transcribes non-coding RNAs that are required for basic functions of the cell, like transfer RNA and spliceosome RNA, but also microRNAs and the transposable element family of short interspersed nuclear element (Dieci et al. 2007).

However, the focus of this thesis is on transcripts produced by RNA polymerase II. The main reason for this is that RNA polymerase II transcribes messenger RNA (mRNA), which is the group of RNA that become translated into protein. The human cell has around 20,000 protein coding genes, all transcribed by RNA polymerase II (Nurk et al. 2022). While RNA polymerase I and III transcribe RNA needed for baseline functions, RNA polymerase II enables the cell to specialize by transcribing only certain protein-coding genes into mRNA.

2.2.2 Transcription is regulated in many steps

Transcription requires many more proteins than RNA polymerase II to properly work. These other proteins are called general transcription factors and are needed to transcribe all mRNA.

Transcription by RNA polymerase II consists of three main phases: Initiation, Elongation and Termination (Lee and Young 2000). Transcription starts by the assembly of the pre-initiation complex. The pre-initiation complex is usually composed of hundreds of

[1] Histone 3 lysine 9

proteins. The minimal pre-initiation complex consists of RNA polymerase II and six of the general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Most of these general transcription factors are themselves protein complexes and can therefore perform multiple required tasks each. There are many additional proteins involved in most cases, especially when enhancers are involved. Additional proteins involved in the pre-initiation complex include chromatin remodelers and the mediator complex (itself consisting of up to 26 subunits) (Allen and Taatjes 2015). The main tasks of the pre-initiation complex are to recruit RNA polymerase II to the transcription start site by recognizing promoter motifs (TFIID, TFIIA) (Ossipow, Fonjallaz, and Schibler 1999), unwind and open the double-stranded DNA to provide access to the DNA template (TFIIE, TFIIF, TFIIH) (Lee and Young 2000) and properly position RNA polymerase II to the active site (TFIIB) (Bushnell et al. 2004). The RNA is then synthesized in a processive manner by RNA polymerase II using the DNA as a template (Kwak and Lis 2013). Before leaving the nucleus, the precursor mRNA goes through multiple post-transcriptional modifications before being exported to the cytoplasm for translation. Transcription is terminated after the recognition of the polyadenylation signal sequence AAUAAA present close to the end of the precursor mRNA (Bienroth, Keller, and Wahle 1993). The 3' end of the precursor mRNA is then cleaved and extended by approximately 250 untemplated adenosine nucleotides. A guanine nucleotide is also attached to the 5' end of the precursor mRNA with a 5' to 5' triphosphate linkage, which promotes nuclear export, translation and intron excision while preventing degradation (Visa et al. 1996; Bird et al. 2016; Konarska, Padgett, and Sharp 1984; Shatkin 1976). The exons are spliced by the spliceosome to generate the final transcript, a process which generate great functional diversity even within the same gene (Marasco and Kornblihtt 2023).

2.3 The genome defines its own regulation

For the different cells in the human (or mammalian) body to be able to specialize in so many kinds of tasks, the regulation of which genes are active at any given time and any given cell must be very precise. With the discovery of mRNA as the physical and informational intermediate between the genetic storage unit of DNA and the functional unit of protein, the regulation of transcription was suggested as a mechanism to control the synthesis of protein (Cobb 2015). The first layer of regulation is whether the gene is physically available at all. Since the genome can be either tightly packed or unpacked by modifying the histone residues present on the nucleosome, the proper conformation of the DNA the gene consist of and the surrounding genomic region, is thought to be important for transcription to occur (Cremer and Cremer 2001). However, there are multiple additional layers of regulation which determine which genes are transcribed.

2.3.1 Transcription factors direct transcription

Transcription factors are DNA-binding proteins that are involved in transcriptional regulation. Many transcription factors function as regulators which determine cell types, drive differentiation and control response pathways. While estimates vary, one review estimated that the human have around 1,700 transcription factor proteins which may be sorted into roughly 70 families (Lambert et al. 2018). Transcription factors work alone, or in a complex, to promote or repress the recruitment of RNA polymerase to specific genes. Transcription factors typically function by either directly recruiting RNA polymerase II or transcriptional cofactors (Frietze and Farnham 2011). Transcription factors recognize relatively short stretches of DNA sequences, known as binding motifs, and bind to either enhancer or promoter elements to effectively decode their instructions. The specificity of a transcription factor to its binding motif is typically multiple orders of magnitude higher compared to noncognate sequences (Geertz, Shore, and Maerkl 2012; Phair et al. 2004). In some cases, this may be the transcription factor's only method of transcriptional regulation: it may simply bind to a motif and block another transcription factor that would promote transcription (Ptashne 2011; Akerblom et al. 1988). However, it is also clear that in eukaryotic genomes the sequence of the binding motif alone is not sufficient. Indeed, most transcription factors only bind some of their target motifs present in the genome, with the rare exception being CTCF (Fu et al. 2008; T. H. Kim et al. 2007). Most motifs are 6-12 base pairs, which do not contain sufficient information content for the transcription factor to specifically bind to sites known to be regulated by that transcription factor (Wunderlich and Mirny 2009). This contrasts with the prokaryotic transcription factor landscape, where the binding motifs typically contain enough information to specifically bind to their intended target. In the eukaryotic case, the paradigm of multiple transcription factors cooperatively and synergistically binding to clusters of binding motifs to direct transcription is a much more favorable theory. Indeed, cis-regulatory regions usually contain many binding sites for multiple different transcription factors, and this fact alone has been used to predict cis-regulatory regions bioinformatically (Berman et al. 2002; Crowley, Roeder, and Bina 1997; Wasserman and Fickett 1998; 1998). Furthermore, the ability for a transcription factor to promote or repress transcription is highly context specific. For example, the same transcription factor can recruit co-factors with opposite effects (Amati and Land 1994). The details on how transcription factors interact with each other biochemically in different configurations are mostly lacking, and there is a current challenge to understand how this complex network of transcription factors work in detail.

2.3.2 Promoters let transcription factors bind close to genes

The core promoter is the DNA segment -40 to +40 base pairs within the transcriptional start site (Roeder 1996). This is the main region where the general transcription factors bind to the DNA. Core promoter elements are DNA motifs that support the assembly of

the pre-initiation complex and directs transcriptional initiation (Haberle and Stark 2018). The core promoter elements can be identified based on their sequence and location in relation to the transcription start site. There is no universally used element present at all genes. Instead, different genes use different core promoter elements to guide the pre-initiation complex. There are broadly two different classes of genes in this respect, genes with a clearly defined core promoter (sharp) and genes with a broad and diffuse core promoter region (broad) (Haberle and Stark 2018). The sharp promoters are generally present in genes that are lineage-specific, while broad promoters are present in genes expressed ubiquitously (Carninci et al. 2006). The core promoter element that was first discovered, the TATA-box, has the consensus sequence 5'-TATAWAW-3' and is located 31 to 24 base pairs upstream of the transcription start site (Goldberg 1979; Bucher 1990). The TATA box is present in about 20% of human genes. Another core promoter, initiator, has the consensus sequence 5'-BBCABW-3' and is present right on top of the transcription start site (Carninci et al. 2006; Vo ngoc et al. 2017). Other core promoters present in human like DRE, TCT, BREu and BREd, all have well defined consensus sequences and positions relative to the transcription start site (Parry et al. 2010; Hirose et al. 1993; W. Deng 2005; Lagrange et al. 1998). The relevance of the DNA sequence of the core promoter downstream of the transcription start site has been unclear in humans. In Drosophila, the downstream DPE and MTE elements are clearly defined by a known location and consensus sequence but matches to these elements in the human genome are rarely observed (Sandelin et al. 2007). A challenge in addressing this seems to have been the inability to use over-representation methods to detect motifs. Recent machine learning approaches have identified an additional downstream core promoter element DPR, which overlap the DPE and MTE elements found in Drosophila (Vo ngoc et al. 2020). Interestingly, the part of the DPR motif that overlap the DPE element are very similar. Furthermore, the presence of DPR is associated with a lack of the TATA box, similar to the situation in Drosophila for the DPE element (Willy, Kobayashi, and Kadonaga 2000).

2.3.3 Enhancers are clusters of transcription factor binding sites

Enhancer elements are regulatory sequences of DNA that are often many kilobases or megabases away from their target gene and are used to activate the transcription of genes in a precise and sensitive manner. One pragmatic definition of enhancer is a cluster of binding sites for transcription factors.

There are multiple different methods to detect enhancer regions in the genome and different to ways to predict whether the enhancer can be considered active in each cell type. By any measure, enhancers are ubiquitous throughout the genome. Based on co-

occurrence of histone modifications H3K27ac[2] and H3K4me1[3] as a marker for enhancers, a large surveying study found 43,011 enhancer candidates across the human genome. Furthermore, the activity of these enhancers were highly cell-type specific (The FANTOM Consortium 2014). A less conservative approach based on DNase I cleavage events found an average of around 330,000 intergenic regions per biosample that may be enhancers (Vierstra et al. 2020). The most recent phenomenon considered to define active enhancer regions is the detection of enhancer RNA, i.e., RNA from transcribed enhancer regions (T.-K. Kim et al. 2010). However, the detection enhancer RNA is difficult due to their short-lived nature and the lack of methods to efficiently detect them (Sartorelli and Lauberth 2020).

While there are different theories on how the enhancer, directly or indirectly, interacts with the promoter region to promote transcription the most favored theory is enhancer-promoter looping (Panigrahi and O'Malley 2021). In this model, the proteins bound by the enhancer and the promoter make physical contact and the protein-protein interactions then further facilitate transcription.

Furthermore, enhancers typically do not interact outside of their topologically associated domain that put limits on which genes they may influence (Cavalheiro, Pollex, and Furlong 2021). Since the genome is a three-dimensional structure, folded and packed in a presumably quite pragmatic fashion, the enhancer might be close to the promoter despite their distant relative position on the linear genome.

2.4 Transcription occurs in bursts

The idea that non-genetic heterogeneity between single cells may arise due to stochastic fluctuations in mRNA molecules was introduced early (Spudich and Koshland 1976); with contemporary observations using electron microscopy which found that the synthesis of multiple RNA molecules is initiated at one time and that there are discrete periods of activity and inactivity of transcription (Miller and Beatty 1969; McKnight and Miller Jr. 1979). A decade later, the observation of transcriptional heterogeneity among single cells in response to glucocorticoid stimulation started the investigation into transcriptional bursts that is still ongoing today (M. S. H. Ko 1991; M. S. Ko, Nakauchi, and Takahashi 1990).

There are many studies examining how the genome, core promoter elements, enhancers and transcription factors influence the bursting behavior of genes. Since transcriptional bursts have been observed throughout the tree of life (Sanchez and Golding 2013), it is likely that the basis for transcriptional bursting, at least in part, relies on some

[2] Acetylation of histone 3 lysine 27
[3] Methylation of histone 3 lysine 4

fundamental aspect of transcription that is consistent across all organisms. The interplay between transcription and torque on the DNA double helix may be a general cause of transcriptional bursting. When RNA polymerase transcribes DNA, torque is applied to the DNA double helix. This causes the DNA behind the RNA polymerase to become less tightly wound and the DNA in front of the polymerase to become more tightly wound. This phenomenon is called positive supercoiling and may be removed by a topoisomerase called DNA gyrase (Liu and Wang 1987). Since the activity of DNA gyrase is limiting, accumulation of positive supercoiling may result in stalling and then result in a burst of transcriptional activity. This mechanism has been shown to be sufficient to explain transcriptional burst in bacteria (Chong et al. 2014) and polymerase spacing has shown to be dependent on torsional stress in eukaryotes (Tantale et al. 2016).

Promoter elements are also involved in shaping transcriptional bursts. One study investigated the actin gene family present in the amoeba *Dictyostelium discoideum* and found that switching the promoters of genes within the family brought with it the transcriptional bursting dynamics for that gene (Tunnacliffe, Corrigan, and Chubb 2018). Due to being the first discovered, the most investigated element out of those is the TATA-box. One study found that the burst size are higher in genes whose core promoter region contain a TATA box and that a mutation in the TATA-box sequence decrease the burst size of the gene (Hornung et al. 2012).

While the exact mechanism is still unclear, most studies suggest enhancer elements direct gene expression changes by affecting the frequency of transcriptional bursts for their target gene. This was first suggested by an early study that examined the effect of the SV40 enhancer on the probability of transcription of a target reporter gene (Walters et al. 1995). Recent publications support this as the most convincing explanation. One study visualized and measured the activity of multiple enhancers in living *Drosophila* embryos and observed dependence of burst frequency on the strength of the activating enhancer element (Fukaya, Lim, and Levine 2016). By forcing the physical contact of the β-globin enhancer to the β-globin gene in mouse cells, another experiment showed that raising the frequency of physical contact increases the burst frequency but not burst size (Bartman et al. 2016). However, physical contact between an enhancer element and its target gene may not be necessary to affect transcription (Alexander et al. 2019).

Transcription factors seem to affect multiple aspects of transcriptional bursts. One paper suggests that burst frequency modulation by the translocation of transcription factors may be a general control strategy to coordinate responses to external stimuli by the study of the transcription factor *Crz1* in yeast. They found that local calcium concentration modulated the frequency of *Crz1* translocation into the nucleus, where each translocation resulted in a multi-gene transcription event (Cai, Dalal, and Elowitz

2008). The similar principle was observed in a modified version of the human cell line MCF7 with a PP7 reporter system for the GREB1 locus, where the concentration of estrogen was shown to modulate the frequency of GREB1 transcription in single cells (Fritzsch et al. 2018). Furthermore, transcription factors are usually a part of signaling pathways that respond to stimuli or give rise to periodical activity. The nature of the signaling pathway dictates the patterns of transcriptional bursts which are observed. Serum induction mediated by the transcription factor serum response factor leads to a transcriptional bursts of the β-actin gene (Kalo et al. 2015). They found a feedback loop where artificially low levels of β-actin leads to an increased transcriptional response. In another study, oscillations in NF-kB localization was shown to control the dynamics of gene expression of its targets by modulating the burst frequency of those genes, including its own negative regulator IkBa (Nelson et al. 2004). This was later also shown for glucocorticoids, a class of steroid hormones that can act as transcription factors, with translocation into the nucleus of the cell as the mechanism (Stavreva et al. 2019). Ultradian patterns in glucocorticoid concentration fluctuations led to transcriptional bursts of the same pattern. Another study also investigated the serum response factor for another gene, c-Fos, with the conclusion that transcription factor concentration modulates the burst frequency (Senecal et al. 2014). Furthermore, they found that the duration of the transcription factor binding event to the DNA binding domain and the strength of the activator domain influenced burst duration (k_{off}) and initiation rate (k_{syn}) respectively. The transcriptional response to DNA damage is regulated by the *p53-Mdm2* system. As a response to DNA damage, the burst frequency of *p53* transcription was reported to increase proportional to the amount of DNA damage. In contrast to the study on c-Fos described above, the burst size and duration was found to be fixed and did not depend of the amount of damage (Lahav et al. 2004).

2.4.1 The impact of transcriptional bursts on phenotype

The initial inquiry into stochastic gene expression was to explain why cells with no genetic differences can be so different, even in the same environment. While there are surely additional factors which may contribute, stochastic gene expression is a clear contributor. One of the more discussed aspects of transcriptional bursting in this regard is its influence on the incomplete penetrance of genetic disorders. Incomplete penetrance is the case where only a fraction of carriers of a mutation develops its associated disease. This phenomenon is thought to be partly caused by stochasticity of gene expression. Transcriptional bursting gives rise to frequent monoallelic expression in diploid systems, which may have an impact on the disease penetrance (Q. Deng et al. 2014). However, examples in literature are scarce. One study used the *Caenorhabditis elegans* model animal to investigate the incomplete penetrance of mutations affecting intestinal specification. They could explain the incomplete penetrance by demonstrating that the variability in gene expression of the mutant alleles altered the topology of the

gene regulatory network that determines intestinal cell fate (Raj et al. 2010). The fluctuations of expression of mutant and wildtype alleles across a tissue may also introduce pathologies. For example, one cause of hypertrophic cardiomyopathy may be due to contractile heterogeneity among the individual cardiomyocytes because of transcriptional heterogeneity of mutant and wild type beta-myosin heavy chain expression among cells (Montag et al. 2018).

In a non-disease setting, variability in gene expression could also be used as a strategy of adaptive evolution, where fluctuations of allele usage within a population may be used as a form of hedging against changes in the immediate environment (Bruijning et al. 2020). From an evolutionary perspective, it might also be beneficial to conserve energy by not transcribing key genes constantly and instead use the RNA template for protein synthesis multiple times, with the precision of gene regulation as a trade-off (Hausser et al. 2019). Transcriptional bursts might also be used to generate diversity during differentiation of stem cells into various committed cell types, which is a concept applicable to most if not all organs in the body. The dynamics of fate commitment of differentiating cells may rely on transcriptional bursts to a significant extent. For example, the lineage commitment of T-cells was recently shown to depend on a stochastic rate-limiting *cis*-epigenetic mechanism at the level of individual gene loci. The activation probability of the gene *Bcl11b* was demonstrated to depend on a distal enhancer region that acts independently for each allele. Furthermore, the transcriptional activation of *Bcl11b* was also dependent on Notch signaling in-*trans* (Ng et al. 2018).

2.5 Allelic expression

Which copy of a gene the cell uses has been the subject of study for a long time. The first studied case was the allelic exclusion of antigen receptors of T and B lymphocytes (Pernis et al. 1965). During lymphocyte maturation, the immunoglobulin gene segments undergo recombination on each allele. Once a productive rearrangement has been achieved on one allele, further rearrangement on the other allele is prevented. In effect, only one allele of the immunoglobulin genes is expressed in mature lymphocytes (Vettermann and Schlissel 2010). There are also examples of genes that are only transcribed from the maternal or paternal allele, these genes are known as imprinted genes (Ferguson-Smith 2011). It has also been suggested that there are autosomal genes that show mitotically heritable monoallelic expression (here called fixed autosomal random monoallelic expression). These genes are suggested to have expression from only one allele in some clones while other clone show biallelic expression. One study based on SNP-sensitive microarrays found that up to 10% of autosomal genes exhibited fixed autosomal random monoallelic expression (Gimelbrant et al. 2007). Other studies showed fixed autosomal random monoallelic expression in mouse lymphoblasts, fibroblasts, human neural stem cells and mouse neural stem cells at similar rates (10%, 2.1%, 1.6-2.2% and 2.4% respectively) (Zwemer et al. 2012; Jeffries et al. 2012; Li et al. 2012). However, observations on the single-cell level contradict some of these findings. Early single cell experiments from the Sandberg lab showed that there is abundant monoallelic expression in the transcriptomes of single cells in mice (Q. Deng et al. 2014). However, most genes did not have a preferred allele to express. Instead, each allele appeared equally often with no obvious pattern. Later experiments on clonally expanded T-cells and fibroblasts showed very few genes that exhibited monoallelic expression which was mitotically stable (<1% of genes) (Reinius et al. 2016). These later experiments used newly developed single-cell sequencing methods, while the prior studies used either bulk RNA-sequencing or microarray technologies. The discrepancies between these findings have been discussed in the literature, and can be attributed to differences between methods and analysis (Vigneau et al. 2018; Reinius and Sandberg 2018).

2.6 X chromosome inactivation and upregulation

Therian mammals, mouse and human included here, have two sex chromosomes, X and Y[4] . Males have both the X and Y chromosome, while females have two X chromosomes. The presence of the Y chromosome, specifically the Y-linked gene SRY, determines the sex during development (Berta et al. 1990). Therefore, the Y chromosome is only ever present in one copy and the X chromosome is present in one copy in males and two copies in females. Out of the two sex chromosomes present in therian mammals, the X chromosome is considerably larger than the Y chromosome, have many more euchromatic regions and more protein-coding genes. They most likely evolved from a regular pair of autosomal chromosomes (Wallis, Waters, and Graves 2008). However, in the present day there is a low amount of homology between the two sex chromosomes. The only homologous areas are the pseudoautosomal regions present at the end of both chromosomes (Helena Mangs and Morris 2007). This allows the two sex chromosomes to pair up during meiosis and are the only areas which may undergo genetic recombination (Ciccodicola et al. 2000). This is presumably the cause of the small size of the Y chromosome, since the X chromosome may recombine in females, but the Y chromosome has lost most genes and degenerated (Graves 2006).

However, this situation creates complications in terms of gene dosage. Importantly, most of the X-linked genes are not involved in sex determination and are required for basic cellular processes (Ross et al. 2005). The female cells effectively have twice the number of X-linked gene copies available to be transcribed and translated compared to males. One way the cell was handling this was detected in 1949, with the report of a "nucleolar satellite" found only in motor neurons of female cats and not male cats, later known as a Barr body (Barr and Bertram 1949). This Barr body was later identified to be a heterochromatic X chromosome (S. Ohno and Hauschka 1960). However, whether this X chromosome was of the paternal or maternal variant was not known. Based on clever observations regarding a number of X-linked mutations that affect the coat color of mice, Mary Lyon suggested that, at least in mice, either of the X chromosomes become heterochromatic early during embryogenesis. Since the heterozygous mutant mice have variegated coats, both the paternal and maternal X chromosome are inactivated in different cells (Lyon 1961). Later, Susumu Ohno hypothesized that the cell have two ways to cope with the unbalanced gene dosage on the X chromosome: one of the X chromosomes is inactivated in females and the genes on the single X chromosome have increased expression (Susumu Ohno 1967). The former is known as X chromosome inactivation, or Lyonization after Mary Lyon. This method of controlling gene dosage of X-linked genes is overwhelmingly established to be a correct theory in eutherian

[4] The monotremes have five pairs of sex chromosomes and don't ask me about monotremes.

mammals (Loda, Collombet, and Heard 2022). However, the same way there are many sex determination strategies in the animal kingdom, the details of this compensation differ between different species. The qualifier eutherian is needed because while marsupials inactivate one of their X chromosomes early in development, it is always the paternal X chromosome (Cooper et al. 1993). Interestingly, studies have observed the same imprinted inactivation of the paternal X chromosome in the mouse around the 4-8 cell stage of the embryo (Okamoto et al. 2004). However, the paternal X is later reactivated, and random X chromosome inactivation take place in the late blastocyst stage. Humans do not seem to have this imprinted inactivation, but rather a transcriptional dampening of both X chromosomes before random X chromosome inactivation that probably takes place in about the same developmental time window as in the mouse (Petropoulos et al. 2016). For both human and mouse, X chromosome inactivation leads to mitotically heritable monoallelic expression for most X-linked genes. A few genes escape X chromosome inactivation and are also transcribed from the inactive X chromosome, albeit in lower abundance.

The second mechanism Susumu Ohno suggested, the upregulation of the single X chromosome (X chromosome upregulation) has been much more debated (Pessia, Engelstädter, and Marais 2014). In contrast to X chromosome inactivation, which according to the 1949 letter to Nature "...may be detected with no more elaborate equipment than a compound microscope following staining of the tissue by the routine Nissl method", X chromosome upregulation first requires the accurate measurement of transcripts from X-linked and autosomal genes. The first analyses based on microarrays and RNA-sequencing data have reached conflicting results (Nguyen and Disteche 2006; Xiong 2010). This disagreement can to a large degree be attributed to what the proper comparison is. First, there is a question of what to compare to in the first place. Susumu Ohno's hypothesis was based on the idea that these genes were originally present on an autosomal chromosome, and the proper comparison should therefore be to the ancestral gene (Susumu Ohno 1967). Most studies have settled on comparing the expression levels of the autosomal genes to the genes on the single active X chromosome. Furthermore, different studies have established different criteria for including genes in their analyses, leading to different conclusions. For example, compared to autosomal chromosomes, the X chromosome contains more genes with little or no gene expression (Kharchenko, Xi, and Park 2011). Other studies claim only some genes on the X chromosome need to be dosage compensated and focused on genes known to be a part of gene networks that include autosomal genes and are therefore presumably more sensitive to gene dosage (Pessia et al. 2012; Lin et al. 2012).

3 Research aims

The aims of this thesis explore the use of transcriptomics to understand transcriptional bursting.

Paper I: Can we use single-cell RNA sequencing data to infer transcriptional bursting parameters genome wide? How is transcriptional bursting encoded in the genome and in which way does it change with cell type and cell state?

Paper II: Do bursts explain monoallelic expression and allelic imbalance?

Paper III: How do burst kinetics change when gene dosage changes on the X chromosome?

Paper IV: Can we use metabolic labelling in single cells to track newly transcribed RNA?

Paper V: What are the advantages to studying newly transcribed RNA in the context for transcriptional bursting?

4 Materials and methods

The studies which form this thesis mainly use mathematical modeling, single-cell RNA sequencing and bioinformatics to study transcriptional bursting. This section briefly describes the methods used in each area.

4.1 Mathematical models for transcriptional bursting

The theory of transcriptional bursting is tightly coupled to its corresponding mathematical models. In one sense, the mathematical models of transcriptional bursting determine the terminology used for describing it.

4.2 The simplest model - the telegraph model

Before describing the model of transcriptional bursting, I will start with the simpler models that may be used to quantitatively describe transcription and why they fail to capture the phenomenon of transcriptional bursts.

The basic deterministic model of transcription includes two parameters: synthesis and degradation. It can be written down as the differential equation

$$\frac{dx}{dt} = \alpha - x\gamma$$

where α is the synthesis rate, γ is the degradation rate and x is the abundance of the RNA. The steady state of this model can easily be found to be the ratio between synthesis rate and the degradation rate, $\frac{\alpha}{\gamma}$. However, this differential equation completely fails to capture any variability in expression observed in single cells. The simplest stochastic model of transcription replaces the two rate parameters above with exponentially distributed waiting times. This gives us the Poisson distribution with the parameter $\lambda = \frac{\alpha}{\gamma}$. The mean stays the same as in the deterministic model while accounting for some variability. Precisely this model was suggested in early studies (Spudich and Koshland 1976). However, because the Poisson model assumes that the RNA molecules are synthesized independently of each other it fails to capture one aspect of variability present in transcriptional bursts, namely the bursts themselves.

The model needs to be extended to account for transcriptional bursts. The simplest model of transcriptional bursts is called the telegraph model. In this model, the gene can either be in an *off* state or an *on* state (Peccoud and Ycart 1995). While in the *on* state the gene is synthesized at rate k_{syn}, and is not transcribed in the *off* state. The time until the system switches from one state to another is described by the exponential probability distribution with an associated k_{off} and k_{on} rate respectively. Each

individual RNA molecule can also be degraded at the rate δ. The steady state distribution is a mixture between a Beta and Poisson distribution. The beta random variable is governed by k_{on} and k_{off} and effectively determines the availability of the gene to be transcribed. The Poisson distribution is determined by k_{syn}. When k_{on} is large, the distribution can be approximated by the Poisson distribution described above with $\frac{k_{syn}}{\delta}$ as the parameter. This is interpretable as the gene always being available for transcription. When both k_{on} and k_{off} are slower than the degradation rate, there are long periods of activity and inactivity. This is observed as a bimodal distribution of transcript abundance. The mean of the distribution is given by the fraction of time the gene is available for transcription times the synthesis rate divided by the degradation rate, $\frac{k_{on}}{k_{on}+k_{off}}\frac{k_{syn}}{\delta}$. The two main units which will be discussed in this thesis are the burst frequency, k_{on}, and burst size, $\frac{k_{syn}}{k_{off}}$, the average number of molecules produced during a burst.

4.2.1 The simplest model can be extended in many ways

While the telegraph model is the most frequently used model, there are extensions of this model which include refractory periods, splicing, cell division, dosage compensation and cell size (Friedman, Cai, and Xie 2006; Shahrezaei and Swain 2008; Stinchcombe, Peskin, and Tranchina 2012; Cao and Grima 2020). Not all these models are analytically tractable, and fewer offer computationally feasible inference methods. Indeed, the main advantage of using the telegraph model is the possibility to infer parameters using snapshot measurements, e.g., single-cell RNA sequencing or single molecule fluorescence in situ hybridization (smFISH). Then applying the moment method, maximum likelihood or Markov-Chain Monte-Carlo (Peccoud and Ycart 1995; Raj et al. 2006; J. K. Kim and Marioni 2013; Gómez-Schiavon et al. 2017; Jiang, Zhang, and Li 2017; Vu et al. 2016). Although this requires the assumptions of stationarity and ergodicity (Dattani and Barahona 2017). Furthermore, the model can also be simplified. If we assume that the rate of gene deactivation is much faster than activation, $k_{off} \gg k_{on}$, this simplified model follows the negative binomial distribution (Shahrezaei and Swain 2008). The trade-off of this simplification is that the model can no longer generate bimodal distributions. Nonetheless, the negative binomial distribution is often used for analyzing single-cell RNA sequencing data (Love, Huber, and Anders 2014; Robinson and Smyth 2007).

4.3 Primary and immortalized cells

To study transcriptional bursts, we need to measure the amount of RNA transcribed for a given gene copy. However, since mammalian genomes are diploid, we need a way to distinguish RNA originating from the two alleles of a gene. Our approach was to use crosses of distantly related subspecies of the *mus musculus* species.

Most often we crossed a female of the CAST/EiJ strain of *mus musculus castaneus* (southeastern Asian house mouse) with a male of the C57BL/6J strain of the *mus musculus domesticus* (western European house mouse) and used the first generation (F1) offspring of this cross. These two mice have many single-nucleotide variants that can be used to distinguish the alleles by sequencing (Keane et al. 2011). One experiment was done using a cell line derived from a F1 CAST/EiJ x 129SvEv cross, but the desired result is the same.

The constituent papers have also used cell lines derived from human cells. The K562 cell line is a myelogenous leukemia cell line derived from a 53 year old female (Lozzio and Lozzio 1975) and the Jurkat T-cell cell line that was derived from a 14 year old male with leukemia (Schwenk and Schneider 1975). As the references report, both cell lines were established in the mid-1970s. However, these cell line do not provide allelic resolution.

4.4 Single-cell sequencing to profile transcription with allelic resolution

Single-cell sequencing enable the relatively unbiased quantification of the polyadenylated transcriptome[5], where polyadenylated transcripts are reverse transcribed and amplified. Since most of the transcribed RNA is ribosomal RNA, it is important to provide a way to select only polyadenylated transcripts. This is done by using an oligo-dT as the primer in the reverse transcription reaction (Mortazavi et al. 2008). After the first method of this kind was reported in 2009 (Tang et al. 2009), the overall field of single-cell sequencing has quickly evolved (Svensson, da Veiga Beltrame, and Pachter 2020). During my PhD studies, the state of the art has changed multiple times when it comes to both wet-lab and computational methods. Furthermore, most of the advances have focused on the development of highly scalable methods followed by shallow sequencing of individual cells to characterize cell types in tissues (Svensson, Vento-Tormo, and Teichmann 2018; Zhang, Ntranos, and Tse 2020). To study transcriptional bursting the experimental focus lies on capturing molecules at high sensitivity and detecting those molecules by sequencing each cell deeply. The most widely used single-cell sequencing protocol, offered by the company 10x Genomics, only captures the 3' end of the transcript (10x Genomics n.d.). For our experimental

[5] Methods which capture non-polyadenylated RNA exist but will not be discussed. Sorry Michael.

approach, this method would be unsuitable, since most single nucleotide polymorphisms which allow us to distinguish the allele of origin are not on the 3' end. Therefore, to capture the allelic expression for most genes, a method which give coverage over the whole transcript is preferred.

4.4.1 The Smart-seq family of methods provide full-length coverage

The Smart-seq family of methods is the most widely used full-length coverage single cell sequencing family of protocols (Ramsköld et al. 2012; Picelli 2013; Hagemann-Jensen et al. 2020; Hagemann-Jensen, Ziegenhain, and Sandberg 2022). The methods rely on a Moloney Murine Leukemia Virus-derived reverse transcriptase (RT) enzyme which often adds 2-5 un-templated nucleotides at the 3'-end of the complementary DNA when the 5'-end of the RNA is reached, where the nucleotide cytosine is preferred (Schmidt and Mueller 1999). This allows the RT enzyme to switch the template to an oligonucleotide that has 3 riboguanosines at its 3'-end (template-switching oligonucleotide). The complementary DNA can then be amplified exponentially using primers targeting sequences present on the oligo-dT and template-switching oligonucleotide sequence. For short-read sequencing, the resulting pool of full-length complementary DNA is then tagmented with the enzyme Tn5 to obtain DNA fragments of a size appropriate for short-read sequencers. Optionally, the tagmentation step can be skipped and DNA library is then sequenced on a long-read sequencer. In the constituent papers, the methods which have been used are Smart-seq3 or a modified version of Smart-seq2 that include a unique molecular identifier in the template-switching oligonucleotide (Hagemann-Jensen et al. 2020). The unique molecular identifier is of particular importance for the estimation of transcriptional bursting estimates since they enable the discrete counting of individual captured RNAs.

4.5 Computational analysis of sequencing data

After sequencing the complementary DNA of the transcripts present in the individual cells, the genes they correspond to need to be identified. This is done through aligning the sequencing reads to a reference genome. The human and mouse genomes have been sequenced and annotated quite extensively compared to most other organisms. Since transcripts are spliced co- and post-transcriptionally, naively aligning reads to the reference genome is not suitable. Therefore, multiple algorithms have been developed that are aware of splicing and can align spliced transcripts (Dobin et al. 2012; D. Kim et al. 2019). More recent methods have been developed that perform pseudoalignment, which instead quantify the compatibility of the sequencing read with a transcript model without performing alignment to the reference genome (Bray et al. 2016; Patro et al. 2017). All the analyses in the constituent papers have used the STAR software for alignment (Dobin et al. 2012). The aligned reads are then stored in the SAM format, which is the standard format for aligned sequencing reads. After alignment, the analyses make

extensive use of a vast number of software libraries to enable further processing, statistical inference, statistical modeling, and visualization (Virtanen et al. 2020; Harris et al. 2020; Hunter 2007). The constituent papers have used the programming languages Python, R and C, with most of the code being written in Python.

4.6 Metabolic labeling

Metabolic labeling is a concept that can be used to study nascent and new transcription of RNAs. A nucleotide analogue is introduced that is partly incorporated during transcription. The incorporation of the nucleotide analogue is then detected in some way to find which RNAs were produced during the metabolic labeling period.

The first sequencing methods using metabolic labeling relied on the physical separation of labelled and unlabeled RNA. In the method TT-seq, thiol-specific biotinylation is followed by affinity purification, and ensures the analyzed complementary DNA originates from newly synthesized RNA (Schwalb et al. 2016).

Single cell applications require very efficient ways to separate the new and old transcriptomes. One recently developed approach moves the separation step from a physical separation to a computational separation after sequencing. By introducing a chemical modification step mutations are introduced into the newly synthesized RNAs, which can be computationally distinguished by a mismatch to the reference genome. The method SLAM-seq was the first method to use this approach. SLAM-seq uses 4-thiouridine followed by an alkylation reaction to induce T-to-C mismatches in reads corresponding to newly synthesized molecules (Herzog et al. 2017). Other methods have been used that adapt this approach to single cells (Erhard et al. 2019). Lastly, another method has been developed that use very efficient click chemistry followed by biotin pull-down to physically separate newly transcribed and old RNA (Battich et al. 2020).

4.7 Ethical considerations

Most of my projects involved sequencing cells from a first-generation offspring of two distantly related strains of Mus musculus, C57BL/6J and CAST. This cross does not lead to any phenotype which the mouse might suffer from. Cells are collected from these mice either post-mortem or from the embryo. The ethical permit under which we did these studies deemed certain activities to be of moderate severity, since some experiments detailed in the permit may require multiple injections of the same animal. However, all the procedures needed for my projects were of mild severity. Nothing is done to the mice that would make them likely to experience any short-term moderate pain, suffering or distress, or long-term mild pain, suffering or distress.

Furthermore, some of my projects are based on already existing data. The analyses in the first submission of Paper I were all done with already existing data, published or otherwise. The questions asked by the reviewers prompted us to generate new data (i.e., use more mice) to answer them. We only did this when it was clear to us that it was necessary to confirm the validity of the results. This data was then used for Paper III which explores other aspects of that dataset, Paper III also uses existing data from another study (Chen et al. 2016). In this regard my projects make large use of the reduce principle of humane animal research, since reusing and reanalyzing data in new ways reduce the number of animals needed and total animal suffering at the same level of scientific output.

5 Results

5.1 Paper I

In Paper I, we wanted to investigate the use of single cell RNA sequencing data to study transcriptional bursting, which first required the development of novel statistical algorithms. To perform statistical inference of transcriptional bursting, I developed a likelihood approach to infer parameters using the two-state model. This estimation strategy differed from previous efforts in multiple ways (Jiang, Zhang, and Li 2017; J. K. Kim and Marioni 2013). The moment likelihood used in (Jiang, Zhang, and Li 2017) had the undesired property of sometimes producing negative rate estimates. Some other advantages were that the profile likelihood technique allowed me to obtain confidence intervals on the point estimates. The likelihood approach also allowed us to compare parameters from two different conditions by comparing their relative likelihoods. While (J. K. Kim and Marioni 2013) used a Markov-Chain Monte-Carlo method that produces more stable estimates than the moment method (although more time consuming), the data we applied our approach to was much more deeply sequenced than previous studies and had an order of magnitude more cells.

I inferred transcriptional burst kinetics from data obtained from the CASTxBL/6J F1 crossbreed for 7,186 genes based on 224 fibroblast cells prepared using a modified version of Smart-seq2. I found the parameters to agree with the previous studies that had been done either on single genes or based on exogenous genes. The burst size ranged between 1-10 RNAs per burst across genes. After scaling the parameters by the degradation rate to obtain the parameters in an absolute time scale, I found that an allele bursts on average every 6 hours. Interestingly, the k_{off} parameter was almost always much larger than k_{on} indicating that genes are often idle with occasional bursts of transcription.

I next investigated the effect of core promoter elements on burst size. I found that the TATA core promoter substantially increases burst size while burst frequency is not affected. This effect is increased by the presence of the Initiator core promoter element, while the Initiator element does not have any effect by itself. No effect was observed based on mean expression or on the level of burst frequency. Furthermore, I found that there was a gene-length dependent effect on burst size, which was not confounded by spliced mRNA length.

We linked enhancer activity to burst frequency regulation using multiple approaches. I compared burst frequency and size of genes expressed in two different cell types: fibroblasts and embryonic stem cells (*n* = 4,854 genes in common). I observed that the main factor that changes between cell types is burst frequency. We performed smFISH on a small number of X-linked genes in male cells. The differences in burst frequency

and size found between the two cell types based on the single cell RNA sequencing data were corroborated by the smFISH data, although the absolute parameter values were somewhat different. Then, I used H3K27ac as a marker of enhancer activity detected by chromatin immunoprecipitation sequencing and used a previously defined enhancer-to-gene map to assign enhancer activity to genes and their corresponding bursting parameters. I compared the relative change in normalized read density over the linked enhancer regions to the relative change in bursting parameters across cell types. The enhancer activity of enhancers linked to genes expressed in both cell types were highly correlated with a corresponding change in burst frequency.

Second, I found that the density of strain specific single nucleotide variants was higher in enhancers of genes with allelic differences in burst frequency compared to genes with similar kinetics.

Last, we inferred transcriptional kinetics in a murine embryonic stem cell line (CAST/EiJ x 129SvEv) with a *Sox2* enhancer deletion on the CAST allele. We found that the resulting reduction in *Sox2* gene expression was due to a reduction in burst frequency. To support this finding, I simulated observations where the corresponding reduction in mean expression was either only due to a reduction in burst frequency or size. The simulations also supported the finding that the enhancer deletion resulted in a reduction in burst frequency.

5.2 Paper II

In Paper II, we wanted to investigate to which extent transcriptional bursts explain monoallelic expression from autosomal genes. The transcriptional burst inference approach from Paper I enabled us to study how transcriptional bursts generate patterns of monoallelic and biallelic expression in single cells. For this paper, we used an expanded dataset of 682 primary fibroblast cells (F1 offspring of CAST/EiJ and C57BL/6J crosses) using the Smart-seq3 method.

I first calculated the theoretical probabilities of observing monoallelic expression according to the two-state model throughout the parameter landscape. When I compared these theoretical values to real data, I found them to agree highly; the observed amount of monoallelic expression closely followed the predicted amount.

The theoretical calculations showed that the amount of monoallelic expression could be modulated by both changes in burst frequency and size. Therefore, I compared the estimated burst frequency and size to the observed fractions of allelic expression and found that burst frequency was the main component of transcriptional bursts that affects monoallelic and biallelic expression. Furthermore, both burst frequency and size contributed to the relative amounts of allelic expression.

I reasoned that cell-type specific gene regulation would lead to an underestimation of biallelic expression if we naively applied the prediction procedure to a heterogeneous group of cells composed of multiple cell types. To investigate this, we sequenced the transcriptomes of single cells derived from the skin of the F1 CASTxC57/BL6 mouse (Smart-seq2, n = 354 cells). We were able to classify the cells into 9 distinct cell types. I first tried to predict the amount of biallelic expression for each gene for the whole population of cells, with the false assumption that all the cells are under identical regulation. I found that genes that are known to be ubiquitously expressed in mouse tissues have biallelic expression consistent with the predicted amount compared to randomly sampled genes, many of which are presumably under cell-type specific regulation. Furthermore, I found that most cell-type clusters have biallelic expression closer to the predicted amount than cells sampled randomly from the whole dataset. By intentionally mixing cell-types with large differences in the transcriptomes, I was able to increase the discrepancy between the predicted and observed amount of biallelic expression (e.g., T-cells and Interfollicular epidermis). However, mixing similar cell types (e.g., Interfollicular epidermis and Lower hair follicle) did not affect this discrepancy.

There have been multiple reports of widespread monoallelic expression and allelic imbalance that are specific to clonal populations of cells. I theoretically investigated if this can be explained solely by variability in expression due to transcriptional bursting. I used the estimated parameter from one of the alleles, CAST, to simulate observations from two alleles *in silico*. Since the underlying parameters are identical, any observed

differences would be due to the statistical variability in the process. I found that for lowly expressed genes, it was common to observe large deviations from equal allelic expression. This effect decreased as the number of simulated observations increased.

5.3 Paper III

In Paper III, we used the data generated in Paper I to investigate the upregulation of the X chromosome in male and female cells in the context of transcriptional bursting, and its relation to X chromosome inactivation in females.

The female cells were classified based on which X chromosome had been deactivated. Then we compared the allelic expression levels of genes on the active X chromosome to the autosomal genes from the same allele. Both the female active X alleles and the male X chromosome consistently showed a higher mean expression as a group compared to autosomal genes by all considered analyses. These analyses included cell-normalized expression levels, differences in the overall mean expression distributions, a sample size-matched permutation test, and pairwise tests between individual chromosomes. This indicated a chromosome-wide upregulation of the X chromosome (X chromosome upregulation).

I then estimated the transcriptional burst kinetic parameters for each group of cells and consistently found higher burst frequencies for the X-linked genes compared to autosomal genes. No burst size differences were detected, and the burst frequency difference was still detected after accounting for differences in degradation.

To investigate whether this upregulation was a fixed or dynamic mechanism, we studied X chromosome upregulation while X chromosome inactivation was taking place. We hypothesized that upregulation was dependent on the number of active X chromosomes. We used a previously published dataset of cells constituting the developmental trajectory from epiblast to the neuronal cell type based on the same F1 crossbreed. During this developmental process, X inactivation take place and the number of active X chromosomes are reduced from two to one. To study the relationship between X chromosome upregulation and X chromosome inactivation, we classified the cells based on stages of X inactivation using X-linked expressed strain-specific variants and measured the extent of X chromosome upregulation within those groups. Before X chromosome inactivation, there is no detectable X chromosome upregulation as measured by a chromosome wide difference in burst frequency. However, as X chromosome inactivation progressed, the active X chromosome showed upregulation by burst frequency proportional to the extent of X chromosome inactivation.

5.4 Paper IV

In Paper IV, we developed a new metabolic labeling method in single cells. The recent advance in metabolic labeling to use sequencing itself as the read out instead of physical separation made it easier to adapt to single cells (Herzog et al. 2017). We developed a single cell sequencing protocol called NASC-seq that uses T>C mismatches to the reference genome, introduced by the nucleotide analogue 4-thiouridine during culturing and converted during RT, to computationally distinguish reads originating from newly transcribed molecules from reads originating from pre-existing molecules. The protocol was developed based on the Smart-seq2 protocol, with the addition of steps specific to metabolic labeling and some other modifications. I implemented a statistical model and data processing pipeline to accurately quantify the proportion of newly transcribed reads while accounting for errors introduced during library preparation and sequencing.

We first applied the NASC-seq protocol to the K562 cell line to demonstrate the successful labeling of newly transcribed RNA, based on the statistical measures (signal-to-noise) and with examples of genes with known high and low turnover. The statistical model was needed to correct measurements from individual cells arising due to errors. We then applied the protocol to Jurkat T-cells stimulated with phorbol myristate acetate and ionomycin while simultaneously exposed to 4-thiouridine. We found that the expected response genes, for example EGR1 and FOS, were exclusive detected as newly transcribed in the stimulated cells. We computationally separated the new and old transcriptomes. The old transcriptomes of stimulated and un-stimulated cells clustered together after dimensionality reduction by principal component analysis, while the new transcriptomes both clustered separately. We identified early response genes from the new transcriptomes of stimulated cells and found this modality to have much better sensitivity to detect the downregulation of genes compared to using the total transcriptome.

I also measured global RNA replacements rates in single unstimulated Jurkat T-cells. At 30 minutes and 60 minutes, a median of 6.5% and 10.8% of the cell transcriptomes had been replaced by new RNA. The replacement rates for individual genes were much more variable compared to cells, where some genes had no RNA replaced and a few had all the RNA replaced. The median replacement was 10.2% and 16.5%, for the 30- and 60-minute labeling conditions respectively.

5.5 Paper V

In Paper V, we developed the next iteration of NASC-seq, NASC-seq2, and applied the method to study transcriptional bursting of newly transcribed RNA.

The developed method, NASC-seq2 is based on the Smart-seq3 protocol with the addition of the metabolic labeling specific steps, which in turn require a dilution step to avoid a high concentration of certain reagents, such as dimethyl sulfoxide, in the downstream reactions which are required for the alkylation to occur. I developed multiple tools to efficiently process NASC-seq2 data. First, I developed a software named *stitcher.py* that is able to reconstruct molecules based on a shared unique molecular identifier (Larsson and Sandberg 2020; Hagemann-Jensen et al. 2020). In contrast to Paper IV, due to the molecule resolution NASC-seq2 provides we can now count new and old RNA molecules instead of the proportion of reads that are new and old. I developed a statistical test using the model from Paper IV to decide which molecules were old and new.

We first applied this method to 613 individual K562 cells and found that NASC-seq2 detects on average 2000 more genes per cells compared to NASC-seq at the same sequencing depth (100,000 total reads). We had a high power to detect individual molecules as new, over 90% power for most genes.

We then applied NASC-seq2 to 8,916 individual primary fibroblasts (C57BL/6 x CAST/EiJ) with 4sU labeling for 2 hours. We detected around 100,000 molecules per cell, with 12.5% of molecules detected as newly transcribed. To infer kinetics based on the newly transcribed RNA, we developed a new inference method that depends on the labeling time and the molecule count distribution. This approach allowed us to investigate k_{off} and k_{syn} separately, which we were unable to do on total RNA measurements. We found that k_{syn} was correlated with burst size but k_{off} was not, indicating that synthesis rates specify the number of RNAs produced in a burst while the window of synthesis stay invariant.

To study co-bursting, we investigate newly transcribed RNA for gene-pairs as a function of their genomic distance. We found that without allelic comparisons, there are many gene-pairs that exhibit positive correlation in new transcription. However, after correcting for correlations seen in non-meaningful *trans* gene and allele pairs (e.g., gene 1, allele 1 compared to gene 2, allele 2), this observed correlation is effectively removed. Among the remaining pairs, there are several paralogues that warrant further investigation, but there is no evidence for genome-wide co-bursting of genes.

6 Discussion

In **Paper I**, we demonstrated the use of single-cell RNA sequencing to infer the kinetics of transcriptional bursting. Furthermore, we discovered several fundamental aspects of transcriptional bursting using this approach.

We were able to further understand how the genome partially encodes for the kinetics of transcriptional bursts. The first example we find is the effect core promoter elements have on burst size. This effect was not detected on the level of mean expression, clearly showing how studying transcriptional bursting give a richer characterization of transcription than studying the average output from single cells. In Paper I, we only reported the effect of two core promoter elements (TATA and Initiator). It is likely a more sophisticated analysis than linear regression will detect the contribution of other core promoter elements to transcriptional bursting than the ones studied here.

The second example was the role of enhancers in dictating the burst frequency of genes. It is important to point out that while the results are convincing on the genome-wide level, the effects of specific enhancer elements or transcription factors on individual genes is much less clear, except for the *Sox2* example. The assignment of enhancers to genes were done with the best methods known at the time and can be improved. In this context, it would be exciting to study how enhancers usage change in the context of cell type and state to affect burst kinetics.

Considering the many steps required to produce an RNA transcript, the emerging picture is that promoters and enhancers affect separate parts of this process. The most straightforward interpretation is that enhancers are involved in forming the pre-initiation complex, while core promoters are mainly involved in later steps that ultimately ensure the successful release of polymerase II to transcribe the gene. This would be the first approximation, but there will surely come more studies which paint a more complicated picture. In particular, the role of proximal promoters is not that clear. Importantly, it is now possible to study these kinds of questions using a genome-wide assay.

Gene regulation, broadly defined as the epigenetic state, is inherited by somatic cells. There are a few established cases of allele level regulation that is somatically inherited, but whether this is a widespread phenomenon shared by most genes is a debated topic. In **Paper II**, we investigated the effect of transcriptional bursting on monoallelic expression and show that monoallelic expression can be fully explained by transcriptional bursts. Furthermore, the main contribution to monoallelic expression is the frequency of bursts. This result argues against any widespread allele level regulation that is stable across cell divisions. On closer inspection, including the results in Figure 5 of Paper II, lowly expressed genes may show allele specific expression which is only due to biological noise in the process of transcription itself. On top of that, technical

variation in the sequencing assay would only exacerbate this problem. Indeed, the genes reported in other studies to exhibit somatically heritable allele regulation were as a group lowly expressed in the studied cell types.

However, monoallelic expression may have effects on cell behavior in shorter time frames than cell generations. Consistent with previous studies, we find that monoallelic expression is abundant. As a result, only one transcript allele is present at any given time in a single cell for many genes. If there are functional differences between the two alleles, the capability of that cell to perform a given task may fluctuate over time. This might not be relevant for genes that perform tasks with high redundancy but would be crucial for genes that perform a very specific task during a limited time window. Transcription factors involved during development are a good example. It is interesting to speculate that genetic disorders that show incomplete penetrance may be explained by the fluctuations of available alleles due to transcriptional bursting.

Another topic Paper II briefly discussed was the possibility to understand homogeneity of a group of cells by exploiting monoallelic expression. Since highly expressed genes in general have a high burst frequency, both alleles tend to be detected. Therefore, detecting a skew in the monoallelic-to-biallelic ratio is evidence of heterogeneity in the regulation of that gene. The results in the paper were convincing, this corollary is certainly reflected in the data, but not immediately useful to apply more broadly. An extension along this line of reasoning would be to develop an explanatory model which explicitly models the biallelic expression due to transcriptional bursting.

The hypothesis that the X chromosome needs to be upregulated after X chromosome inactivation to compensate for haplo-insufficiency and maintain fitness was first presented by Susumu Ohno in 1967. This idea has been investigated using multiple different methods and the results have been conflicting. In **Paper III**, we showed how genome-wide transcriptional burst inference allows us to detect chromosome-level regulation of transcriptional bursting. Interestingly, we showed that the X chromosome upregulation is responsive to the number of active X chromosomes in the cell. Furthermore, the change in gene expression is driven by changes in the frequency of transcriptional bursts. This shows the capability of the cell to detect gene dosage conditions and adapt the regulation of transcriptional bursting.

The measured upregulation was around 1.4x, which is considerably lower than the output of two active alleles (2x). It is plausible that while 1.4x does not fully compensate for the inactive allele, the overall up-regulation is sufficient to ensure the gene products are present to the extent needed by the cell. Another explanation may be the involvement of multiple other modes of upregulation, for example in translation or degradation of the RNA transcripts. However, exactly how the cell would increase translation or decrease degradation of transcripts that are specifically X-linked is unclear.

How the transcriptional bursting on the X chromosome is modulated to achieve higher burst frequencies is the next obvious question. In the paper we hypothesize based on Paper I that enhancers are involved in dynamically changing the burst frequency. As the inactivated X progressively becomes inaccessible and heterochromatic, it is likely the local concentration of *trans*-acting factors shifts to the active X. This is further supported by the observation that X-linked genes that escape inactivation are expressed at a lower burst frequency from the inactive X compared to the active X. Indeed, follow-up studies confirmed that the degree of upregulation is tightly linked to the degree of inactivation.

It is possible the modulation of transcriptional bursting is a general adaptive strategy to compensate for any aneuploidy. The most interesting applications for this direction would be trisomy 21 (Down's syndrome) and copy-number variation or chromosomal aberrations often observed in cancer.

In **Paper IV**, we developed a method to detect newly transcribed RNA in single cells. This method was greatly improved in **Paper V**, where we studied the transcriptional burst kinetics of newly transcribed RNA in single cells.

In Paper IV, we used metabolic labelling to measure the new transcriptomes of cells after the exposure of a perturbation (phorbol myristate acetate and ionomycin). Only the new transcriptome contained the actual perturbation response, while the old transcriptome was similar to the transcriptome of the control cells. This experiment demonstrated the power of using metabolic labelling to study transcriptional responses to perturbations, in particular the ability to study only the transcriptome after perturbation. We found that many responses, especially downregulation of genes, were hidden on the total transcriptome level but clearly visible on the new transcriptome level. Responses of genes that are already transcribed but increase in expression were also more readily detected. Using metabolic labelling for this purpose gives a more accurate and richer characterization of transcriptional responses to perturbations and should in my opinion be the standard. Furthermore, a future study could use the improved NASC-seq2 to study the burst kinetics of these kind of transcriptional responses.

In the time between Paper IV and Paper V, the amount of data we could collect exploded. The main accelerators of this increase in throughput were the capacity of short-read sequencing and the extensive work in the wetlab to automate protocols (Hagemann-Jensen, Ziegenhain, and Sandberg 2022). I must say I was responsible for neither of these. However, it did require a massive improvement of the performance of multiple algorithms and software components in the analysis workflow, as they could not scale to the new requirements.

The improved NASC-seq2 protocol together with a new inference framework allowed us to study transcriptional bursting genome-wide on a more detailed level. The time-

resolved nature of the metabolic labelling data enabled us to investigate k_{off} and k_{syn} separately instead of studying burst size (k_{syn}/k_{off}). We found that all genes with inferable kinetics exhibited transcriptional bursting, where the duration of bursting is relatively constant, but the rate of synthesis is variable. Based on this result and other studies, I argue that genes are all transcribed in bursts and that there is no such thing as "bursty genes" and "non-bursty genes", since there is no evidence in the literature there is any another mode of transcription.

The idea that genes are transcribed simultaneously is usually discussed in the context of transcriptional hubs. There are countless of articles where this concept is discussed. Genes needed for the same task tends to be localized near each other on the chromosome and active in the same cell type. However, there has been no clear evidence that genes tend to be transcribed together. NASC-seq2 can show that this is not the case in general. Furthermore, we show that allelic resolution is needed for this analysis, since extrinsic factors may introduce correlations on the gene level. However, paralogues in proximity seem to be a special case. Since paralogues share an ancestral gene, they are often regulated by the same factors which would enable co-regulation. The coordination of transcription in time could be of large significance for certain cellular tasks and it is possible the relative chromosomal location of the paralogues that show co-bursting are under selective pressure specifically to facilitate this phenomenon.

One drawback of single cell RNA sequencing is its inherently destructive nature. To subject the cell to the protocol it must be lysed. Therefore, it is not possible to probe the cell at a future time point since we killed it. Metabolic labelling is a strategy to recover some time-resolved information even though the transcriptome can only be sampled once per cell. However, this approach does have some limitations. To provide the metabolic label, the cells typically must be in culture. The label could in theory be administered *in vivo* to an animal model or organoids, although that approach would require extensive validation that the label is successfully delivered to all target cells and in equal concentration. This clearly restricts the kind of experiments one could perform, for example studies on primary human cells are not possible. Nevertheless, metabolic labelling in single cells is an excellent approach to study fundamental aspects of transcription and transcriptional responses to perturbations.

7 Conclusions

In conclusion, my thesis has explored the possibility to measure transcriptional burst kinetics with single cell transcriptomics, which led to many insights into mechanisms of transcription in mammalian cells. The techniques are still improving, and it will be exciting to further discover more underlying principles of transcriptional bursting.

8 Acknowledgements

Dear Reader, welcome to the most important section of the thesis. This is probably the first, and possibly only, section you will read of this book. I'm saying this because that's what I do most of the time. It takes a village to get through a PhD, no matter what, and the people you're surrounded by become very important.

The first person to acknowledge is my main supervisor **Rickard.** Thank you for the opportunity to do by PhD in your lab and for all the support along the way. In particular, the freedom to pursue many different questions and projects. But also, for how you helped me manage through the more difficult periods of my PhD during the pandemic and afterwards.

I would also like to acknowledge my two co-supervisors, **Yudi** and **Yishao**. Although you both did not end up that involved in my PhD studies, you both helped me gain fundamental skills to successfully complete my PhD.

Next, I would like to say thank you to everyone in the **Sandberg lab**. Especially in the context of this thesis all your feedback on the text and your remarkable turn-around time due to my last-minute request for help. I was not surprised my request would end up being last-minute nor that you all would be there to help in time.

More specifically, in a particular order but please do not read too much into it, I'd like to thank **Leo.** I think the speech that I held at your graduation party said most of what I would like to say here. In the abridged version, I'd like to thank you for all the adventures we've had together and all the support. Excited to see what you do next! **Gert-Jan**, thank you for our extensive collaboration on the constitutive papers. You've always been available for discussions, feedback, and mentoring. **Michael,** thanks for all the hugs. I probably needed those. Also thank you for the science and the mentoring. Your work ethic has been inspiring. Thank you, **Daniel**, for the work together and your deep theoretical knowledge that I think enrich the whole lab. Thank you, **Christoph**, for all the fun rants about academia. **Jens,** your PhD thesis was inspiring for me while doing my own PhD, and I'd like to thank you for that. Thank you, **Juliane**, and good luck on the remainder of your PhD. Thank you **Gösta** for being the lab guru. Thanks to the recent students in the lab, **Paloma, Salomé, Gustav** and **Arnold**.

Of course, there have been members of the lab which have since left. Thank you **Per**, for your crucial smFISH, abundance of Dill Chips, and our shared support for Liverpool FC. YNWA. Best of luck at Astra. **Björn,** thank you for teaching me about the X chromosome and how interesting it is, and for your mentoring. You've been doing great in your own lab, and I am excited to see what more will come. **Lisa,** thank you for the collaborations. Hope you're doing well in Germany! Thanks to the medical student **Oscar,** which helped with computational work. Thanks to **Åsa** and **Omid** for your work on Paper I. Thank you

Ping for being a good desk neighbor for some time. Thank you, **Emma**, for your lab admin which was always helpful to me. Thanks to the two previous PhD students, **Mtakai** and **Ilgar**, for being role models during the early part of my PhD studies.

I'd also like to acknowledge people from the **Frisén lab**, past and present. Although none of the constitutive papers include any of them, they've been the lab's neighbours for the last five years. **Ionut**, thank you for all the dinners and parties and conversations (this also goes to **Johannes**). **Enric,** for your inquisitive mind and shared love of food. **Giuseppe** (and **Alexandra**)**,** thanks for the great conversations about science, life and Dungeons and Dragons! Thanks to other lab members: **Margherita, Johanna, Moa, Camilla, Carl-Johan, Qirong, Martyna, Mathew, Helena, Jeff, Marta and Embla** for lunch, conversation and lots of other things. And thanks to **Jonas** as well!

To the former members of the Bergmann lab, **Marion** and **Enikö,** you were both great lab neighbours and friends (you're both still friends but you used to be too). Even though I'm not a member of your book club I'm always curious to know what's next.

Thanks to all the teachers at the master's programme where I taught for four years, especially thank you **Lars-Arne.** I learned a lot while teaching myself. Also, thank you **Linda** and **Matti.**

I'd also like to thank all the co-authors that haven't yet been mentioned: **Chloe, Bing, Tina, Tim, Maria, Christos, Michael, Katja** and **Patrick**.

I would now like to acknowledge some friends.

Bobby! Even though I think we both gave each other terrible first impressions, those didn't matter. You've been a great friend since we met, and I've always enjoyed your stories. I have great memories from my trip to Bulgaria and would love to visit again. I'm happy to see you graduate too and returning to medicine. Thank you, **Nathan,** for all the fun we've had together. I'm looking forward to seeing you graduate a couple of days after me. **Lucas.** We have spent an unhealthy amount of time playing World of Warcraft together. But it was worth it because you're a great guy to hang out with. Let me know if you want to try the hardcore version that they're releasing soon. I heard you can play as a duo. Thanks **Kevin,** for crypto news, Svelte community drama and being a cool dude. **Adam,** thank you and good luck on your own defense later this year! **Oscar,** it's been a long time since we first met during my bachelor's studies. Thank you for many philosophical conversations. I look forward to your graduation as well! **Richard,** thanks for being a supportive and kind friend ever since high school. **Emelie** and **Dominyka**, thank you both. Thank you, **Sara.** For anyone that knows me and decided to look up my thesis to read it but hasn't been mentioned, thank **You**.

Tack **Mamma,** för att du uppfostrade mig till den person jag är idag. Lät mig vara nyfiken och undra fritt. Tack till min syster **Agnes**.

Finally, **Ilke**, there are many reasons to thank you. I think most of them are better said at our wedding. I'd like to thank you for all your support all these years. There have been ups and downs during my PhD, and you've been there to celebrate the successes and motivated me in harder moments. I can't express enough how important you've been. Seni seviyorum.

9 References

10x Genomics. n.d. “Single Cell Gene Expression.” 10x Genomics. Accessed May 17, 2023. https://www.10xgenomics.com/products/single-cell-gene-expression.

Akerblom, Ingrid E., Emily P. Slater, Miguel Beato, John D. Baxter, and Pamela L. Mellon. 1988. “Negative Regulation by Glucocorticoids Through Interference with a CAMP Responsive Enhancer.” *Science* 241 (4863): 350–53. https://doi.org/10.1126/science.2838908.

Alexander, Jeffrey M, Juan Guan, Bingkun Li, Lenka Maliskova, Michael Song, Yin Shen, Bo Huang, Stavros Lomvardas, and Orion D Weiner. 2019. “Live-Cell Imaging Reveals Enhancer-Dependent Sox2 Transcription in the Absence of Enhancer Proximity.” *ELife* 8 (May): e41769. https://doi.org/10.7554/eLife.41769.

Allen, Benjamin L., and Dylan J. Taatjes. 2015. “The Mediator Complex: A Central Integrator of Transcription.” *Nature Reviews Molecular Cell Biology* 16 (3): 155–66. https://doi.org/10.1038/nrm3951.

Amati, Bruno, and Hartmut Land. 1994. “Myc—Max—Mad: A Transcription Factor Network Controlling Cell Cycle Progression, Differentiation and Death.” *Current Opinion in Genetics & Development* 4 (1): 102–8. https://doi.org/10.1016/0959-437X(94)90098-1.

Babu, Arvind, and Ram S. Verma. 1987. “Chromosome Structure: Euchromatin and Heterochromatin.” In *International Review of Cytology*, edited by G. H. Bourne, K. W. Jeon, and M. Friedlander, 108:1–60. Academic Press. https://doi.org/10.1016/S0074-7696(08)61435-7.

Barr, Murray L., and Ewart G. Bertram. 1949. “A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis.” *Nature* 163 (4148): 676–77. https://doi.org/10.1038/163676a0.

Bartman, Caroline R., Sarah C. Hsu, Chris C.-S. Hsiung, Arjun Raj, and Gerd A. Blobel. 2016. “Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping.” *Molecular Cell* 62 (2): 237–47. https://doi.org/10.1016/j.molcel.2016.03.007.

Battich, Nico, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S. Baron, Marvin E. Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. 2020. “Sequencing Metabolically Labeled Transcripts in Single Cells Reveals MRNA Turnover Strategies.” *Science* 367 (6482): 1151–56. https://doi.org/10.1126/science.aax3072.

Berman, Benjamin P., Yutaka Nibu, Barret D. Pfeiffer, Pavel Tomancak, Susan E. Celniker, Michael Levine, Gerald M. Rubin, and Michael B. Eisen. 2002. “Exploiting Transcription Factor Binding Site Clustering to Identify Cis-Regulatory Modules Involved in Pattern Formation in the Drosophila Genome.” *Proceedings of the National Academy of Sciences* 99 (2): 757–62. https://doi.org/10.1073/pnas.231608898.

Berta, Philippe, J. Boss Hawkins, Andrew H. Sinclair, Anne Taylor, Beatrice L. Griffiths, Peter N. Goodfellow, and Marc Fellous. 1990. “Genetic Evidence Equating SRY and the

Testis-Determining Factor." *Nature* 348 (6300): 448–50. https://doi.org/10.1038/348448a0.

Bienroth, S, W Keller, and E Wahle. 1993. "Assembly of a Processive Messenger RNA Polyadenylation Complex." *The EMBO Journal* 12 (2): 585–94.

Bird, Jeremy G., Yu Zhang, Yuan Tian, Natalya Panova, Ivan Barvík, Landon Greene, Min Liu, et al. 2016. "The Mechanism of RNA 5′ Capping with NAD+, NADH, and Desphospho-CoA." *Nature* 535 (7612): 444–47. https://doi.org/10.1038/nature18622.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Bruijning, Marjolein, C. Jessica E. Metcalf, Eelke Jongejans, and Julien F. Ayroles. 2020. "The Evolution of Variance Control." *Trends in Ecology & Evolution* 35 (1): 22–33. https://doi.org/10.1016/j.tree.2019.08.005.

Bucher, Philipp. 1990. "Weight Matrix Descriptions of Four Eukaryotic RNA Polymerase II Promoter Elements Derived from 502 Unrelated Promoter Sequences." *Journal of Molecular Biology* 212 (4): 563–78. https://doi.org/10.1016/0022-2836(90)90223-9.

Bushnell, David A., Kenneth D. Westover, Ralph E. Davis, and Roger D. Kornberg. 2004. "Structural Basis of Transcription: An RNA Polymerase II-TFIIB Cocrystal at 4.5 Angstroms." *Science* 303 (5660): 983–88. https://doi.org/10.1126/science.1090838.

Cai, Long, Chiraj K. Dalal, and Michael B. Elowitz. 2008. "Frequency-Modulated Nuclear Localization Bursts Coordinate Gene Regulation." *Nature* 455 (7212): 485–90. https://doi.org/10.1038/nature07292.

Cao, Zhixing, and Ramon Grima. 2020. "Analytical Distributions for Detailed Models of Stochastic Gene Expression in Eukaryotic Cells." *Proceedings of the National Academy of Sciences* 117 (9): 4682–92. https://doi.org/10.1073/pnas.1910888117.

Carninci, Piero, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, et al. 2006. "Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution." *Nature Genetics* 38 (6): 626–35. https://doi.org/10.1038/ng1789.

Cavalheiro, Gabriel R, Tim Pollex, and Eileen EM Furlong. 2021. "To Loop or Not to Loop: What Is the Role of TADs in Enhancer Function and Gene Regulation?" *Current Opinion in Genetics & Development*, Genome Architecture and Expression, 67 (April): 119–29. https://doi.org/10.1016/j.gde.2020.12.015.

Chen, Geng, John Paul Schell, Julio Aguila Benitez, Sophie Petropoulos, Marlene Yilmaz, Björn Reinius, Zhanna Alekseenko, et al. 2016. "Single-Cell Analyses of X Chromosome Inactivation Dynamics and Pluripotency during Differentiation." *Genome Research* 26 (10): 1342–54. https://doi.org/10.1101/gr.201954.115.

Chong, Shasha, Chongyi Chen, Hao Ge, and X. Sunney Xie. 2014. "Mechanism of Transcriptional Bursting in Bacteria." *Cell* 158 (2): 314–26. https://doi.org/10.1016/j.cell.2014.05.038.

Ciccodicola, Alfredo, Maurizio D'Esposito, Teresa Esposito, Fernando Gianfrancesco, Carmela Migliaccio, Maria Giuseppina Miano, Maria Rosaria Matarazzo, et al. 2000. "Differentially Regulated and Evolved Genes in the Fully Sequenced Xq/Yq Pseudoautosomal Region." *Human Molecular Genetics* 9 (3): 395–401. https://doi.org/10.1093/hmg/9.3.395.

Cobb, Matthew. 2015. "Who Discovered Messenger RNA?" *Current Biology* 25 (13): R526–32. https://doi.org/10.1016/j.cub.2015.05.032.

Cooper, D. W., P. G. Johnston, J. M. Watson, and J. A. M. Graves. 1993. "X-Inactivation in Marsupials and Monotremes." *Seminars in Developmental Biology* 4 (2): 117–28. https://doi.org/10.1006/sedb.1993.1014.

Cremer, T., and C. Cremer. 2001. "Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells." *Nature Reviews Genetics* 2 (4): 292–301. https://doi.org/10.1038/35066075.

Crowley, Evelyn M, Kathryn Roeder, and Minou Bina. 1997. "A Statistical Model for Locating Regulatory Regions in Genomic DNA." *Journal of Molecular Biology* 268 (1): 8–14. https://doi.org/10.1006/jmbi.1997.0965.

Dattani, Justine, and Mauricio Barahona. 2017. "Stochastic Models of Gene Transcription with Upstream Drives: Exact Solution and Sample Path Characterization." *Journal of The Royal Society Interface* 14 (126): 20160833. https://doi.org/10.1098/rsif.2016.0833.

Deng, Q., D. Ramsköld, B. Reinius, and R. Sandberg. 2014. "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells." *Science* 343. https://doi.org/10.1126/science.1245316.

Deng, W. 2005. "A Core Promoter Element Downstream of the TATA Box That Is Recognized by TFIIB." *Genes & Development* 19 (20): 2418–23. https://doi.org/10.1101/gad.342405.

Dieci, Giorgio, Gloria Fiorino, Manuele Castelnuovo, Martin Teichmann, and Aldo Pagano. 2007. "The Expanding RNA Polymerase III Transcriptome." *Trends in Genetics* 23 (12): 614–22. https://doi.org/10.1016/j.tig.2007.09.001.

Dixon, Jesse R., David U. Gorkin, and Bing Ren. 2016. "Chromatin Domains: The Unit of Chromosome Organization." *Molecular Cell* 62 (5): 668–80. https://doi.org/10.1016/j.molcel.2016.05.018.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Erhard, Florian, Marisa A. P. Baptista, Tobias Krammer, Thomas Hennig, Marius Lange, Panagiota Arampatzi, Christopher S. Jürges, Fabian J. Theis, Antoine-Emmanuel Saliba, and Lars Dölken. 2019. "ScSLAM-Seq Reveals Core Features of Transcription Dynamics in Single Cells." *Nature* 571 (7765): 419–23. https://doi.org/10.1038/s41586-019-1369-y.

Ferguson-Smith, Anne C. 2011. "Genomic Imprinting: The Emergence of an Epigenetic Paradigm." *Nature Reviews. Genetics* 12 (8): 565–75. https://doi.org/10.1038/nrg3032.

Friedman, Nir, Long Cai, and X. Sunney Xie. 2006. "Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression." *Physical Review Letters* 97 (16): 168302. https://doi.org/10.1103/PhysRevLett.97.168302.

Frietze, Seth, and Peggy J. Farnham. 2011. "Transcription Factor Effector Domains." In *A Handbook of Transcription Factors*, edited by Timothy R. Hughes, 261–77. Subcellular Biochemistry. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-9069-0_12.

Fritzsch, Christoph, Stephan Baumgärtner, Monika Kuban, Daria Steinshorn, George Reid, and Stefan Legewie. 2018. "Estrogen-dependent Control and Cell-to-cell Variability of Transcriptional Bursting." *Molecular Systems Biology* 14 (2): e7678. https://doi.org/10.15252/msb.20177678.

Fu, Yutao, Manisha Sinha, Craig L. Peterson, and Zhiping Weng. 2008. "The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome." *PLOS Genetics* 4 (7): e1000138. https://doi.org/10.1371/journal.pgen.1000138.

Fukaya, Takashi, Bomyi Lim, and Michael Levine. 2016. "Enhancer Control of Transcriptional Bursting." *Cell* 166 (2): 358–68. https://doi.org/10.1016/j.cell.2016.05.025.

Geertz, Marcel, David Shore, and Sebastian J. Maerkl. 2012. "Massively Parallel Measurements of Molecular Interaction Kinetics on a Microfluidic Platform." *Proceedings of the National Academy of Sciences* 109 (41): 16540–45. https://doi.org/10.1073/pnas.1206011109.

Gimelbrant, Alexander, John N. Hutchinson, Benjamin R. Thompson, and Andrew Chess. 2007. "Widespread Monoallelic Expression on Human Autosomes." *Science (New York, N.Y.)* 318 (5853): 1136–40. https://doi.org/10.1126/science.1148910.

Goldberg, M.L. 1979. *Sequence Analysis of Drosophila Histone Genes*. Thesis. Stanford University.

Gómez-Schiavon, Mariana, Liang-Fu Chen, Anne E. West, and Nicolas E. Buchler. 2017. "BayFish: Bayesian Inference of Transcription Dynamics from Population Snapshots of Single-Molecule RNA FISH in Single Cells." *Genome Biology* 18 (1). https://doi.org/10.1186/s13059-017-1297-9.

Haberle, Vanja, and Alexander Stark. 2018. "Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation." *Nature Reviews Molecular Cell Biology* 19 (10): 621–37. https://doi.org/10.1038/s41580-018-0028-8.

Hagemann-Jensen, Michael, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton J. M. Larsson, Omid R. Faridani, and Rickard Sandberg. 2020. "Single-Cell RNA Counting at Allele and Isoform Resolution Using Smart-Seq3." *Nature Biotechnology* 38 (6): 708–14. https://doi.org/10.1038/s41587-020-0497-0.

Hagemann-Jensen, Michael, Christoph Ziegenhain, and Rickard Sandberg. 2022. "Scalable Single-Cell RNA Sequencing from Full Transcripts with Smart-Seq3xpress." *Nature Biotechnology* 40 (10): 1452–57. https://doi.org/10.1038/s41587-022-01311-4.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hausser, Jean, Avi Mayo, Leeat Keren, and Uri Alon. 2019. "Central Dogma Rates and the Trade-off between Precision and Economy in Gene Expression." *Nature Communications* 10 (1): 68. https://doi.org/10.1038/s41467-018-07391-8.

Helena Mangs, A, and Brian J Morris. 2007. "The Human Pseudoautosomal Region (PAR): Origin, Function and Future." *Current Genomics* 8 (2): 129–36.

Herzog, Veronika A., Brian Reichholf, Tobias Neumann, Philipp Rescheneder, Pooja Bhat, Thomas R. Burkard, Wiebke Wlotzka, Arndt von Haeseler, Johannes Zuber, and Stefan L. Ameres. 2017. "Thiol-Linked Alkylation of RNA to Assess Expression Dynamics." *Nature Methods* 14 (12): 1198–1204. https://doi.org/10.1038/nmeth.4435.

Hirose, F., M. Yamaguchi, H. Handa, Y. Inomata, and A. Matsukage. 1993. "Novel 8-Base Pair Sequence (Drosophila DNA Replication-Related Element) and Specific Binding Factor Involved in the Expression of Drosophila Genes for DNA Polymerase Alpha and Proliferating Cell Nuclear Antigen." *The Journal of Biological Chemistry* 268 (3): 2092–99.

Hornung, G., R. Bar-Ziv, D. Rosin, N. Tokuriki, D. S. Tawfik, M. Oren, and N. Barkai. 2012. "Noise-Mean Relationship in Mutated Promoters." *Genome Research* 22 (12): 2409–17. https://doi.org/10.1101/gr.139378.112.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

International Human Genome Sequencing Consortium. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

Jeffries, Aaron R., Leo W. Perfect, Julia Ledderose, Leonard C. Schalkwyk, Nicholas J. Bray, Jonathan Mill, and Jack Price. 2012. "Stochastic Choice of Allelic Expression in Human Neural Stem Cells." *Stem Cells (Dayton, Ohio)* 30 (9): 1938–47. https://doi.org/10.1002/stem.1155.

Jiang, Yuchao, Nancy R. Zhang, and Mingyao Li. 2017. "SCALE: Modeling Allele-Specific Gene Expression by Single-Cell RNA Sequencing." *Genome Biology* 18 (1): 74. https://doi.org/10.1186/s13059-017-1200-8.

Kalo, Alon, Itamar Kanter, Amit Shraga, Jonathan Sheinberger, Hadar Tzemach, Noa Kinor, Robert H. Singer, Timothée Lionnet, and Yaron Shav-Tal. 2015. "Cellular Levels of Signaling Factors Are Sensed by β-Actin Alleles to Modulate Transcriptional Pulse Intensity." *Cell Reports* 11 (3): 419–32. https://doi.org/10.1016/j.celrep.2015.03.039.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. "Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation." *Nature* 477 (7364): 289–94. https://doi.org/10.1038/nature10413.

Kharchenko, Peter V., Ruibin Xi, and Peter J. Park. 2011. "Evidence for Dosage Compensation between the X Chromosome and Autosomes in Mammals." *Nature Genetics* 43 (12): 1167–69. https://doi.org/10.1038/ng.991.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

Kim, Jong Kyoung, and John C Marioni. 2013. "Inferring the Kinetics of Stochastic Gene Expression from Single-Cell RNA-Sequencing Data." *Genome Biology* 14 (1): r7. https://doi.org/10.1186/gb-2013-14-1-r7.

Kim, Tae Hoon, Ziedulla K. Abdullaev, Andrew D. Smith, Keith A. Ching, Dmitri I. Loukinov, Roland D. Green, Michael Q. Zhang, Victor V. Lobanenkov, and Bing Ren. 2007. "Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome." *Cell* 128 (6): 1231–45. https://doi.org/10.1016/j.cell.2006.12.048.

Kim, Tae-Kyung, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, et al. 2010. "Widespread Transcription at Neuronal Activity-Regulated Enhancers." *Nature* 465 (7295): 182–87. https://doi.org/10.1038/nature09033.

Ko, M. S., H. Nakauchi, and N. Takahashi. 1990. "The Dose Dependence of Glucocorticoid-Inducible Gene Expression Results from Changes in the Number of Transcriptionally Active Templates." *The EMBO Journal* 9 (9): 2835–42. https://doi.org/10.1002/j.1460-2075.1990.tb07472.x.

Ko, Minoru S.H. 1991. "A Stochastic Model for Gene Induction." *Journal of Theoretical Biology* 153 (2): 181–94. https://doi.org/10.1016/S0022-5193(05)80421-7.

Konarska, Maria M., Richard A. Padgett, and Phillip A. Sharp. 1984. "Recognition of Cap Structure in Splicing in Vitro of MRNA Precursors." *Cell* 38 (3): 731–36. https://doi.org/10.1016/0092-8674(84)90268-X.

Kwak, Hojoong, and John T. Lis. 2013. "Control of Transcriptional Elongation." *Annual Review of Genetics* 47: 483–508. https://doi.org/10.1146/annurev-genet-110711-155440.

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg, and R. H. Ebright. 1998. "New Core Promoter Element in RNA Polymerase II-Dependent Transcription: Sequence-Specific DNA Binding by Transcription Factor IIB." *Genes & Development* 12 (1): 34–44. https://doi.org/10.1101/gad.12.1.34.

Lahav, Galit, Nitzan Rosenfeld, Alex Sigal, Naama Geva-Zatorsky, Arnold J Levine, Michael B Elowitz, and Uri Alon. 2004. "Dynamics of the P53-Mdm2 Feedback Loop in Individual Cells." *Nature Genetics* 36 (2): 147–50. https://doi.org/10.1038/ng1293.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. "The Human Transcription Factors." *Cell* 172 (4): 650–65. https://doi.org/10.1016/j.cell.2018.01.029.

Larsson, Anton JM, and Rickard Sandberg. 2020. "Stitcher.Py." Zenodo. https://doi.org/10.5281/zenodo.3765223.

Lee, T. I., and R. A. Young. 2000. "Transcription of Eukaryotic Protein-Coding Genes." *Annual Review of Genetics* 34: 77–137. https://doi.org/10.1146/annurev.genet.34.1.77.

Li, Sierra M., Zuzana Valo, Jinhui Wang, Hanlin Gao, Chauncey W. Bowers, and Judith Singer-Sam. 2012. "Transcriptome-Wide Survey of Mouse CNS-Derived Cells Reveals Monoallelic Expression within Novel Gene Families." *PloS One* 7 (2): e31751. https://doi.org/10.1371/journal.pone.0031751.

Lin, Fangqin, Ke Xing, Jianzhi Zhang, and Xionglei He. 2012. "Expression Reduction in Mammalian X Chromosome Evolution Refutes Ohno's Hypothesis of Dosage Compensation." *Proceedings of the National Academy of Sciences* 109 (29): 11752–57. https://doi.org/10.1073/pnas.1201816109.

Liu, L. F., and J. C. Wang. 1987. "Supercoiling of the DNA Template during Transcription." *Proceedings of the National Academy of Sciences* 84 (20): 7024–27. https://doi.org/10.1073/pnas.84.20.7024.

Loda, Agnese, Samuel Collombet, and Edith Heard. 2022. "Gene Regulation in Time and Space during X-Chromosome Inactivation." *Nature Reviews Molecular Cell Biology* 23 (4): 231–49. https://doi.org/10.1038/s41580-021-00438-7.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lozzio, CB, and BB Lozzio. 1975. "Human Chronic Myelogenous Leukemia Cell-Line with Positive Philadelphia Chromosome." *Blood* 45 (3): 321–34. https://doi.org/10.1182/blood.V45.3.321.321.

Luger, Karolin, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. 1997. "Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution." *Nature* 389 (6648): 251–60. https://doi.org/10.1038/38444.

Lyon, M. F. 1961. "Gene Action in X-Chromosome of the Mouse (Mus Musculus L.)." *Nature* 190. https://doi.org/10.1038/190372a0.

Marasco, Luciano E., and Alberto R. Kornblihtt. 2023. "The Physiology of Alternative Splicing." *Nature Reviews Molecular Cell Biology* 24 (4): 242–54. https://doi.org/10.1038/s41580-022-00545-z.

McKnight, Steven L., and Oscar L. Miller Jr. 1979. "Post-Replicative Nonribosomal Transcription Units in D. Melanogaster Embryos." *Cell* 17 (3): 551–63. https://doi.org/10.1016/0092-8674(79)90263-0.

Miller, O. L., and B. R. Beatty. 1969. "Visualization of Nucleolar Genes." *Science* 164 (3882): 955–57. https://doi.org/10.1126/science.164.3882.955.

Montag, Judith, Kathrin Kowalski, Mirza Makul, Pia Ernstberger, Ante Radocaj, Julia Beck, Edgar Becker, et al. 2018. "Burst-Like Transcription of Mutant and Wildtype MYH7-Alleles as Possible Origin of Cell-to-Cell Contractile Imbalance in Hypertrophic Cardiomyopathy." *Frontiers in Physiology* 9: 359. https://doi.org/10.3389/fphys.2018.00359.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Nelson, D. E., A. E. C. Ihekwaba, M. Elliott, J. R. Johnson, C. A. Gibney, B. E. Foreman, G. Nelson, et al. 2004. "Oscillations in NF-KappaB Signaling Control the Dynamics of

Gene Expression." *Science (New York, N.Y.)* 306 (5696): 704–8. https://doi.org/10.1126/science.1099962.

Ng, Kenneth KH, Mary A Yui, Arnav Mehta, Sharmayne Siu, Blythe Irwin, Shirley Pease, Satoshi Hirose, Michael B Elowitz, Ellen V Rothenberg, and Hao Yuan Kueh. 2018. "A Stochastic Epigenetic Switch Controls the Dynamics of T-Cell Lineage Commitment." *ELife* 7: e37851. https://doi.org/10.7554/elife.37851.

Nguyen, D. K., and C. M. Disteche. 2006. "Dosage Compensation of the Active X Chromosome in Mammals." *Nat. Genet* 38. https://doi.org/10.1038/ng1705.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. "The Complete Sequence of a Human Genome." *Science* 376 (6588): 44–53. https://doi.org/10.1126/science.abj6987.

Ohno, S., and T. S. Hauschka. 1960. "Allocycly of the X-Chromosome in Tumors and Normal Tissues*." *Cancer Research* 20 (4): 541–45.

Ohno, Susumu. 1967. *Sex Chromosomes and Sex-Linked Genes*. Springer.

Okamoto, Ikuhiro, Arie P. Otte, C. David Allis, Danny Reinberg, and Edith Heard. 2004. "Epigenetic Dynamics of Imprinted X Inactivation During Early Mouse Development." *Science* 303 (5658): 644–49. https://doi.org/10.1126/science.1092727.

Ossipow, Vincent, Philippe Fonjallaz, and Ueli Schibler. 1999. "An RNA Polymerase II Complex Containing All Essential Initiation Factors Binds to the Activation Domain of PAR Leucine Zipper Transcription Factor Thyroid Embryonic Factor." *Molecular and Cellular Biology* 19 (2): 1242–50.

Panigrahi, Anil, and Bert W. O'Malley. 2021. "Mechanisms of Enhancer Action: The Known and the Unknown." *Genome Biology* 22 (1): 108. https://doi.org/10.1186/s13059-021-02322-1.

Parry, T. J., J. W. M. Theisen, J.-Y. Hsu, Y.-L. Wang, D. L. Corcoran, M. Eustice, U. Ohler, and J. T. Kadonaga. 2010. "The TCT Motif, a Key Component of an RNA Polymerase II Transcription System for the Translational Machinery." *Genes & Development* 24 (18): 2013–18. https://doi.org/10.1101/gad.1951110.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Peccoud, J., and B. Ycart. 1995. "Markovian Modeling of Gene-Product Synthesis." *Theoretical Population Biology* 48 (2): 222–34. https://doi.org/10.1006/tpbi.1995.1027.

Pernis, B., G. Chiappino, A. S. Kelus, and P. G. Gell. 1965. "Cellular Localization of Immunoglobulins with Different Allotypic Specificities in Rabbit Lymphoid Tissues." *The Journal of Experimental Medicine* 122 (5): 853–76. https://doi.org/10.1084/jem.122.5.853.

Pessia, Eugénie, Jan Engelstädter, and Gabriel A. B. Marais. 2014. "The Evolution of X Chromosome Inactivation in Mammals: The Demise of Ohno's Hypothesis?" *Cellular and Molecular Life Sciences* 71 (8): 1383–94. https://doi.org/10.1007/s00018-013-1499-6.

Pessia, Eugénie, Takashi Makino, Marc Bailly-Bechet, Aoife McLysaght, and Gabriel A. B. Marais. 2012. "Mammalian X Chromosome Inactivation Evolved as a Dosage-Compensation Mechanism for Dosage-Sensitive Genes on the X Chromosome." *Proceedings of the National Academy of Sciences* 109 (14): 5346–51. https://doi.org/10.1073/pnas.1116763109.

Petropoulos, Sophie, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. "Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos." *Cell* 165 (4): 1012–26. https://doi.org/10.1016/j.cell.2016.03.023.

Phair, Robert D., Paola Scaffidi, Cem Elbi, Jaromíra Vecerová, Anup Dey, Keiko Ozato, David T. Brown, Gordon Hager, Michael Bustin, and Tom Misteli. 2004. "Global Nature of Dynamic Protein-Chromatin Interactions In Vivo: Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins." *Molecular and Cellular Biology* 24 (14): 6393–6402. https://doi.org/10.1128/MCB.24.14.6393-6402.2004.

Picelli, S. 2013. "Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nat. Methods* 10. https://doi.org/10.1038/nmeth.2639.

Ptashne, Mark. 2011. "Principles of a Switch." *Nature Chemical Biology* 7 (8): 484–87. https://doi.org/10.1038/nchembio.611.

Raj, Arjun, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. 2006. "Stochastic MRNA Synthesis in Mammalian Cells." *PLoS Biology* 4 (10): e309. https://doi.org/10.1371/journal.pbio.0040309.

Raj, Arjun, Scott A. Rifkin, Erik Andersen, and Alexander van Oudenaarden. 2010. "Variability in Gene Expression Underlies Incomplete Penetrance." *Nature* 463 (7283): 913–18. https://doi.org/10.1038/nature08781.

Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R. Faridani, Gregory A. Daniels, et al. 2012. "Full-Length MRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells." *Nature Biotechnology* 30 (8): 777–82. https://doi.org/10.1038/nbt.2282.

Reinius, Björn, Jeff E Mold, Daniel Ramsköld, Qiaolin Deng, Per Johnsson, Jakob Michaëlsson, Jonas Frisén, and Rickard Sandberg. 2016. "Analysis of Allelic Expression Patterns in Clonal Somatic Cells by Single-Cell RNA–Seq." *Nature Genetics* 48 (11): 1430–35. https://doi.org/10.1038/ng.3678.

Reinius, Björn, and Rickard Sandberg. 2018. "Reply to 'High Prevalence of Clonal Monoallelic Expression.'" *Nature Genetics* 50 (9): 1199–1200. https://doi.org/10.1038/s41588-018-0189-6.

Robinson, Mark D., and Gordon K. Smyth. 2007. "Moderated Statistical Tests for Assessing Differences in Tag Abundance." *Bioinformatics* 23 (21): 2881–87. https://doi.org/10.1093/bioinformatics/btm453.

Roeder, Robert G. 1996. "The Role of General Initiation Factors in Transcription by RNA Polymerase II." *Trends in Biochemical Sciences* 21 (9): 327–35. https://doi.org/10.1016/S0968-0004(96)10050-5.

Rosenfeld, Jeffrey A, Zhibin Wang, Dustin E Schones, Keji Zhao, Rob DeSalle, and Michael Q Zhang. 2009. "Determination of Enriched Histone Modifications in Non-Genic Portions of the Human Genome." *BMC Genomics* 10 (March): 143. https://doi.org/10.1186/1471-2164-10-143.

Ross, Mark T., Darren V. Grafham, Alison J. Coffey, Steven Scherer, Kirsten McLay, Donna Muzny, Matthias Platzer, et al. 2005. "The DNA Sequence of the Human X Chromosome." *Nature* 434 (7031): 325–37. https://doi.org/10.1038/nature03440.

Russell, Jackie, and Joost C.B.M. Zomerdijk. 2006. "The RNA Polymerase I Transcription Machinery." Edited by Stefan G.E. Roberts, Robert O.J. Weinzierl, and Robert J. White. *Biochemical Society Symposia* 73 (January): 203–16. https://doi.org/10.1042/bss0730203.

Saksouk, Nehmé, Elisabeth Simboeck, and Jérôme Déjardin. 2015. "Constitutive Heterochromatin Formation and Transcription in Mammals." *Epigenetics & Chromatin* 8 (1): 3. https://doi.org/10.1186/1756-8935-8-3.

Sanchez, A., and I. Golding. 2013. "Genetic Determinants and Cellular Constraints in Noisy Gene Expression." *Science* 342 (6163): 1188–93. https://doi.org/10.1126/science.1242975.

Sandelin, Albin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David A. Hume. 2007. "Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies." *Nature Reviews Genetics* 8 (6): 424–36. https://doi.org/10.1038/nrg2026.

Sartorelli, Vittorio, and Shannon M. Lauberth. 2020. "Enhancer RNAs Are an Important Regulatory Layer of the Epigenome." *Nature Structural & Molecular Biology* 27 (6): 521. https://doi.org/10.1038/s41594-020-0446-0.

Schmidt, Wolfgang M., and Manfred W. Mueller. 1999. "CapSelect: A Highly Sensitive Method for 5′ CAP-Dependent Enrichment of Full-Length CDNA in PCR-Mediated Analysis of MRNAs." *Nucleic Acids Research* 27 (21): e31–i. https://doi.org/10.1093/nar/27.21.e31-i.

Schwalb, Björn, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. 2016. "TT-Seq Maps the Human Transient Transcriptome." *Science* 352 (6290): 1225–28. https://doi.org/10.1126/science.aad9841.

Schwenk, Hans-Ulrich, and Ulrich Schneider. 1975. "Cell Cycle Dependency of a T-Cell Marker on Lymphoblasts." *Blut: Zeitschrift Für Die Gesamte Blutforschung* 31 (5): 299–306. https://doi.org/10.1007/BF01634146.

Senecal, Adrien, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E. Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. 2014. "Transcription Factors Modulate C-Fos Transcriptional Bursts." *Cell Reports* 8 (1): 75–83. https://doi.org/10.1016/j.celrep.2014.05.053.

Shahrezaei, V., and P. S. Swain. 2008. "Analytical Distributions for Stochastic Gene Expression." *Proceedings of the National Academy of Sciences* 105 (45): 17256–61. https://doi.org/10.1073/pnas.0803850105.

Shatkin, A. J. 1976. "Capping of Eucaryotic MRNAs." *Cell* 9 (4): 645–53. https://doi.org/10.1016/0092-8674(76)90128-8.

Spudich, John L., and D. E. Koshland. 1976. “Non-Genetic Individuality: Chance in the Single Cell.” *Nature* 262 (5568): 467–71. https://doi.org/10.1038/262467a0.

Stavreva, Diana A., David A. Garcia, Gregory Fettweis, Prabhakar R. Gudla, George F. Zaki, Vikas Soni, Andrew McGowan, et al. 2019. “Transcriptional Bursting and Co-Bursting Regulation by Steroid Hormone Release Pattern and Transcription Factor Mobility.” *Molecular Cell* 75 (6): 1161–1177.e11. https://doi.org/10.1016/j.molcel.2019.06.042.

Stinchcombe, Adam R., Charles S. Peskin, and Daniel Tranchina. 2012. “Population Density Approach for Discrete MRNA Distributions in Generalized Switching Models for Stochastic Gene Expression.” *Physical Review E* 85 (6): 061919. https://doi.org/10.1103/PhysRevE.85.061919.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter. 2020. “A Curated Database Reveals Trends in Single-Cell Transcriptomics.” *Database* 2020 (January): baaa073. https://doi.org/10.1093/database/baaa073.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann. 2018. “Exponential Scaling of Single-Cell RNA-Seq in the Past Decade.” *Nature Protocols* 13 (4): 599–604. https://doi.org/10.1038/nprot.2017.149.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. “MRNA-Seq Whole-Transcriptome Analysis of a Single Cell.” *Nature Methods* 6 (5): 377–82. https://doi.org/10.1038/nmeth.1315.

Tantale, Katjana, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, et al. 2016. “A Single-Molecule View of Transcription Reveals Convoys of RNA Polymerases and Multi-Scale Bursting.” *Nature Communications* 7 (1). https://doi.org/10.1038/ncomms12248.

The FANTOM Consortium, Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, et al. 2014. “An Atlas of Active Enhancers across Human Cell Types and Tissues.” *Nature* 507 (7493): 455–61. https://doi.org/10.1038/nature12787.

Tunnacliffe, Edward, Adam M. Corrigan, and Jonathan R. Chubb. 2018. “Promoter-Mediated Diversification of Transcriptional Bursting Dynamics Following Gene Duplication.” *Proceedings of the National Academy of Sciences* 115 (33): 8364–69. https://doi.org/10.1073/pnas.1800943115.

Vettermann, Christian, and Mark S. Schlissel. 2010. “Allelic Exclusion of Immunoglobulin Genes: Models and Mechanisms.” *Immunological Reviews* 237 (1): 22–42. https://doi.org/10.1111/j.1600-065X.2010.00935.x.

Vierstra, Jeff, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, et al. 2020. “Global Reference Mapping of Human Transcription Factor Footprints.” *Nature* 583 (7818): 729–36. https://doi.org/10.1038/s41586-020-2528-x.

Vigneau, Sébastien, Svetlana Vinogradova, Virginia Savova, and Alexander Gimelbrant. 2018. “High Prevalence of Clonal Monoallelic Expression.” *Nature Genetics* 50 (9): 1198–99. https://doi.org/10.1038/s41588-018-0188-7.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for

Scientific Computing in Python." *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Visa, N., E. Izaurralde, J. Ferreira, B. Daneholt, and I. W. Mattaj. 1996. "A Nuclear Cap-Binding Complex Binds Balbiani Ring Pre-MRNA Cotranscriptionally and Accompanies the Ribonucleoprotein Particle during Nuclear Export." *The Journal of Cell Biology* 133 (1): 5–14. https://doi.org/10.1083/jcb.133.1.5.

Vo ngoc, Long, California Jack Cassidy, Cassidy Yunjing Huang, Sascha H.C. Duttke, and James T. Kadonaga. 2017. "The Human Initiator Is a Distinct and Abundant Element That Is Precisely Positioned in Focused Core Promoters." *Genes & Development* 31 (1): 6–11. https://doi.org/10.1101/gad.293837.116.

Vo ngoc, Long, Cassidy Yunjing Huang, California Jack Cassidy, Claudia Medrano, and James T. Kadonaga. 2020. "Identification of the Human DPR Core Promoter Element Using Machine Learning." *Nature* 585 (7825): 459–63. https://doi.org/10.1038/s41586-020-2689-7.

Volpe, Thomas A., Catherine Kidner, Ira M. Hall, Grace Teng, Shiv I. S. Grewal, and Robert A. Martienssen. 2002. "Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi." *Science* 297 (5588): 1833–37. https://doi.org/10.1126/science.1074973.

Vu, Trung Nghia, Quin F. Wills, Krishna R. Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. 2016. "Beta-Poisson Model for Single-Cell RNA-Seq Data Analyses." *Bioinformatics* 32 (14): 2128–35. https://doi.org/10.1093/bioinformatics/btw202.

Wallis, M. C., P. D. Waters, and J. A. M. Graves. 2008. "Sex Determination in Mammals—before and after the Evolution of SRY." *Cell. Mol. Life Sci.* 65. https://doi.org/10.1007/s00018-008-8109-z.

Walters, M. C., S. Fiering, J. Eidemiller, W. Magis, M. Groudine, and D. I. Martin. 1995. "Enhancers Increase the Probability but Not the Level of Gene Expression." *Proceedings of the National Academy of Sciences* 92 (15): 7125–29. https://doi.org/10.1073/pnas.92.15.7125.

Wasserman, Wyeth W, and James W Fickett. 1998. "Identification of Regulatory Regions Which Confer Muscle-Specific Gene Expression11Edited by G. Von Heijne." *Journal of Molecular Biology* 278 (1): 167–81. https://doi.org/10.1006/jmbi.1998.1700.

Willy, P. J., R. Kobayashi, and J. T. Kadonaga. 2000. "A Basal Transcription Factor That Activates or Represses Transcription." *Science (New York, N.Y.)* 290 (5493): 982–85. https://doi.org/10.1126/science.290.5493.982.

Wunderlich, Zeba, and Leonid A. Mirny. 2009. "Different Gene Regulation Strategies Revealed by Analysis of Binding Motifs." *Trends in Genetics* 25 (10): 434–40. https://doi.org/10.1016/j.tig.2009.08.003.

Xiong, Y. 2010. "RNA Sequencing Shows No Dosage Compensation of the Active X-Chromosome." *Nat. Genet.* 42. https://doi.org/10.1038/ng.711.

Zhang, Martin Jinye, Vasilis Ntranos, and David Tse. 2020. "Determining Sequencing Depth in a Single-Cell RNA-Seq Experiment." *Nature Communications* 11 (1): 774. https://doi.org/10.1038/s41467-020-14482-y.

Zwemer, Lillian M, Alexander Zak, Benjamin R Thompson, Andrew Kirby, Mark J Daly, Andrew Chess, and Alexander A Gimelbrant. 2012. "Autosomal Monoallelic Expression in the Mouse." *Genome Biology* 13 (2): R10. https://doi.org/10.1186/gb-2012-13-2-r10.