Model-driven analysis of gene expression control

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Jérémie Breda

von Schweiz, Genève GE

2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

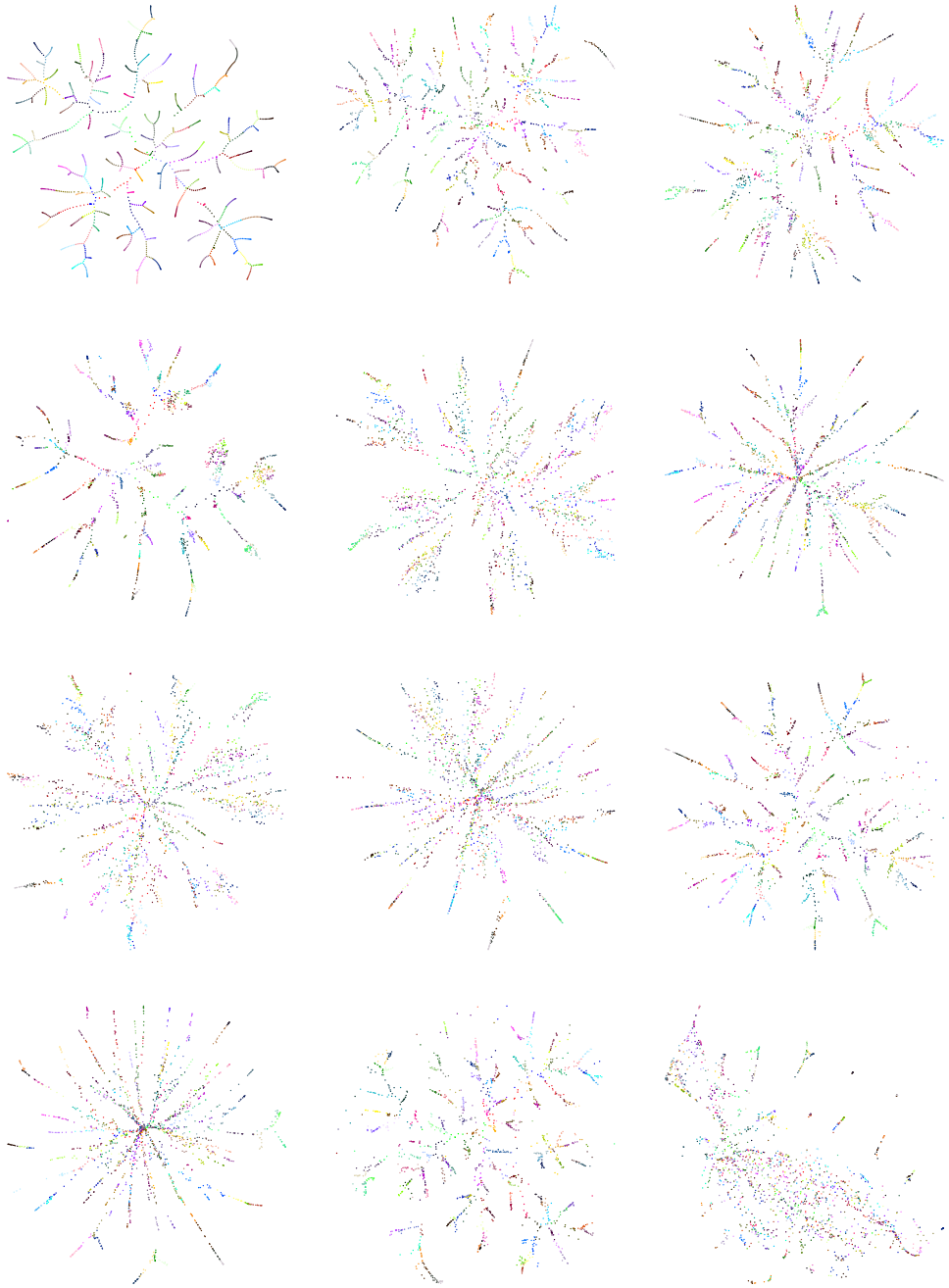
Prof. Dr. Erik van Nimwegen (Faculty representative)

Prof. Dr. Felix Naef (Co-referee)

Basel, 18. Februar 2020

Prof. Dr. Martin Spiess
Dekan

Model-driven analysis of gene expression control



Jérémie Breda

Acknowledgment

This Thesis has been made possible thank to many, and I would like here to thank the people who contributed to the successful completion of my PhD.

Firstly, I would like to express my sincere gratitude to my supervisors Mihaela Zavolan and Erik van Nimwegen for their dedication and support throughout my entire PhD journey, they truly went beyond and above to make a better scientist out of me, and always found time out of their busy schedule when I was in need. Mihaela has been an inspiring mentor, always on top of every project running in the lab, supporting individual ideas and initiative while keeping a focused vision towards successful outcomes. Erik has been a great mentor whose passion and pedagogy never stopped inspiring me, helped me develop myself as a scientist, and kept my motivation high all along my doctoral studies.

I had the chance to be supervised by Andrzej Rzepiela during my master thesis and the beginning of my doctoral studies. I greatly thank him for setting me on such good tracks from the beginning of my PhD, which helped me all the way to defending it.

Besides my supervisors, I would like to thank the members of my thesis committee Felix Naef and Richard Neher. They brought valuable advises and encouragements as well as critical perspectives that pushed me to always question what I take for granted.

I gratefully thank all my scientific collaborators who were at the heart of my research, abundantly shared their knowledge and expertise, and with whom I exchanged insightful and stimulating discussions.

I would also like to thank the Werner-Siemens Foundation and the University of Basel for awarding me a Werner-Siemens Fellowship for Excellence. This provided me independent funding and included me in an interdisciplinary network of fellows that widen my research from various perspectives.

This work was entirely computational and would never have been possible without the high-performance computing services and infrastructures of the University of Basel. All members of the scicore team have been exceptionally supportive and professional, and the computing environment they provided and maintained directly contributed to all my research during these doctoral studies.

Many thanks to Yvonne Steger, Sarah Güthe, and Rita Manohar for their amazing administrative and personal support throughout the years of my PhD.

I had the chance of spending my time at the Biozentrum among exceptional fellow labmates and very good friends. I want to sincerely thank them for all the stimulating discussions and the uncountable memorable moments of fun we shared together.

I want to warmly thank my parents who invariably supported me and advised me all along my studies and on many other aspects of my live. Thanks also to my sister who is always there when I need help, advise or support.

Special thanks to Charlotte for being such a great friend and partner in rising our child. She never stopped supporting me and made it possible to combine my role as a father together with my studies, and I want to sincerely thank her for that.

Finally, I want to thank my wife Andrea. All along those years, she always gave her most sincere advice, never stopped encouraging me to overcome my weaknesses, and has been an unconditional support during the best and the worse times. Most importantly, she contributed to fill my time in Basel with unforgettable memories; we shared countless cheerful occasion, joyful moments and exciting times.

To my beloved girls Lilly and Eleonore,
their dedication and resilience,
their kindness and joyfulness.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multicellular eukaryotes possess up to trillions of cells, all deriving through cell divisions from one fertilised egg. As the DNA polymerase only introduces about one mutation in hundred thousand nucleotides, and the proofreading and DNA repair mechanisms further correct 99% of the replication errors, the mutation rate during cell division is brought down to 10^{-10} to 10^{-8} [Pray, 2008]. The genome size ranging from 10^7 to 10^{10} [Maloy and Hughes, 2013], it follows that the number of mutations occurring per genome per division is between 10^{-3} and 10^2, (0.3 for humans and 20 for mice). These numbers highlight that all cells of a multicellular organism have essentially identical genetic information, the same DNA. Strikingly however, this information is translated into an astonishing variety of cells, which in humans is thought to amount to hundreds of different cell types. These differ from one another by the level of expression of different genomic regions. This is accomplished by regulatory mechanisms that control the usage of DNA information, and act at different steps of gene expression, from chromatin remodelling, to transcription, RNA splicing and processing, export to cytoplasm, translation into proteins and ultimately degradation [Orphanides and Reinberg, 2002].

1.1 Regulation of transcription

DNA is tightly packed around nucleosomes (composed of histone proteins) and folded in a complex 3-dimensional structure, known as chromatin. This packing makes a large fraction of the genome relatively inaccessible to transcription by the RNA polymerase, thus offering opportunity for regulation [Clapier and Cairns, 2009]. Remodelling factors are recruited to reposition and eject nucleosomes, as well as unwrap the DNA, making it accessible for transcription in a condition-dependent manner. Four major, highly conserved families of remodelers contribute to the sliding, ejecting and spacing of nucleosomes, to promote or repress gene expression. Remodelers target individual histones by recognising enzymatically-generated chemical modifications (methylation, phosphorylation, acetylation, ubiquitylation, citrullination, sumoylation) of specific amino acids [Clapier and Cairns, 2009, Phillips, 2008]. The methylation of the DNA itself can also influence directly its accessibility to trans-acting factors [Clapier and Cairns, 2009]. DNA and histone modifications critically impact the accessibility of cis-elements that are located in promoter regions, near transcription start sites. These elements bind transcription factors (TFs), some of which act as "master" regulators of gene expression [Lee and Young, 2013]. TFs are DNA-binding proteins that activate gene expression by recruiting the RNA polymerase or repress gene expression by occluding the transcription start site. With likely more than 1600 TFs and more than 20000 genes in the human genome [Lambert et al., 2018], each gene can be regulated in a unique manner, through a unique combination of TF binding sites. Work in the past decade has also uncovered a large number of another type of cis-regulatory element, enhancers [Schoenfelder and Fraser, 2019]. Enhancers are in many ways similar to promoters, containing TF binding sites and supporting regulatory interactions between specific TFs and gene regulatory elements [Pennacchio et al., 2013]. However, enhancers differ from promoters in their location, being located more distally from the regulated gene, up to 1Mbp upstream and

downstream [Pennacchio et al., 2013], compared to promoters, which are thought to reside within a kilobase of the transcription start site [Bussemaker et al., 2001, Thompson, 2003, Bajic et al., 2004, Carninci et al., 2006]. It has long been presumed that enhancers and promoters interact [Kulaeva et al., 2012], though the mechanism behind this interaction was unknown. Breakthrough research in the past few years has led to the view that complexes of cohesin proteins create loops around the DNA, bringing regions that are far apart in the linear sequence in close proximity, supporting their interaction with the same regulatory proteins [Shlyueva et al., 2014, Schoenfelder and Fraser, 2019, Davidson et al., 2019].

1.2 Post-transcriptional regulation

According to the gene and transcript annotation available in the *Ensembl genome database* (www.ensembl.org), the average number of transcripts for a protein-coding gene varies from 1 in yeast and up to more than 7 in human, with an average around 2 transcripts per gene for most species among the 212 that are annotated (shows Fig. 1.1a). In each distinct species, the number of transcripts per protein coding gene follows an exponential distribution, most genes having a small number of transcripts and some genes having a very large number of transcripts (Fig. 1.1b).

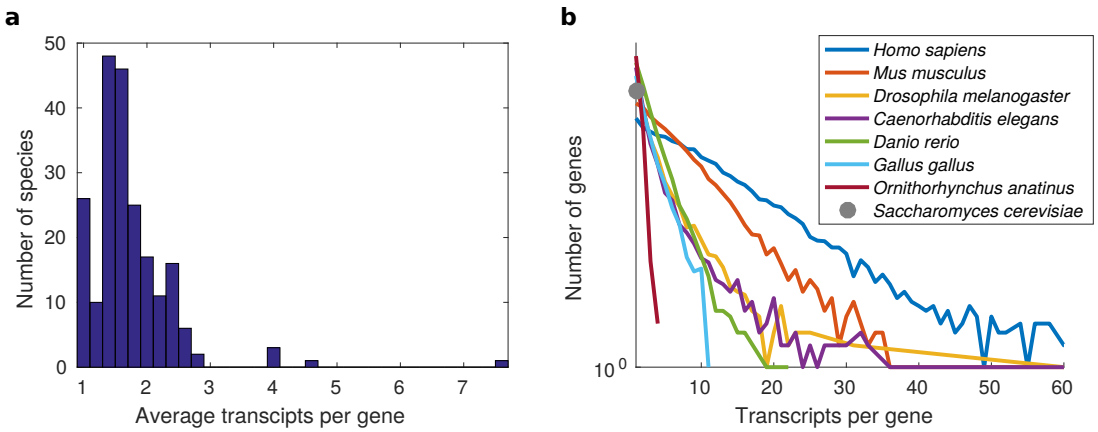


Figure 1.1: **Statistics of number of transcripts per gene.** (**a**) Histogram of the average number of transcripts per protein coding gene computed across the genome annotation of 212 species. **b** Distribution of transcripts per protein coding gene across 8 different species, as indicated on the legend.

The production of distinct isoforms is regulated through the use of alternative promoters [Landry et al., 2003], alternative splicing and alternative termination sites [Smith and Valcárcel, 2000, Lee and Rio, 2015]. It has been estimated that 52% of human genes are regulated by alternative promoter usage [Kimura et al., 2006], and 28% of mouse genes [Tsuritani et al., 2007]. While alternative promoter usage does not necessarily imply changes in the open reading frame [Zavolan et al., 2002, Landry et al., 2003], the

promoter structure can affect the interaction between transcription and splicing factors splicing [Kornblihtt, 2005], thereby influencing downstream splicing. The exonic composition of the mature mRNA is determined by the splicing of the pre-mRNA by the spliceosome, a large complex of proteins and small nuclear RNAs that forms stepwise on the pre-mRNAs guided by signals known as splice site. The relative 'strength' of splice sites as well as additional cis-acting regulatory elements that function as binding sites for trans-acting factors lead to enhanced or reduced splicing at specific exons, and thereby production of alternative splicing isoforms [Wahl et al., 2009]. The secondary structure of the mRNA and indirect effects of the chromatin state also affect alternative splicing [Lee and Rio, 2015]. The 3' end of most mammalian transcripts are determined by the process of 3' end cleavage and polyadenylation (reviewed in [Gruber and Zavolan, 2019]). The process is again guided by short sequence elements that are recognised by specific factors that cleave the pre-mRNA at specific sites and add a poly(A) tail, a sequence composed almost exclusively of adenosines reaching up to 250 nucleotides [Gruber and Zavolan, 2019]. In protein-coding transcripts, the coding region only covers a fraction of the transcript, being flanked by 5' and 3' untranslated regions (UTRs). 5' UTRs have typical role in controlling translation, while the composition of 3'-UTRs influences the stability, localisation and translation rate of mRNAs; in mammalian genes, 3' UTRs are comparable in length to the coding region and have a large impact on the processing and function of the mRNA [Gruber and Zavolan, 2019]. Cleavage and polyadenylation of the 3' ends is undertaken by large complex of factors binding specific sequence motifs of the pre-mRNA [Mandel et al., 2008], and alternative choice of the poly(A) site usage, a specific AAUAAA motif, depends on the expression of regulatory RNA-binding proteins [Gruber and Zavolan, 2019]. The transport of the mature mRNA from the nucleus to the cytoplasm is supported by RNA-binding proteins such as splicing factors and poly(A)-binding proteins that act as adaptors and allow the nuclear export receptor heterodimer to transport mature mRNAs through the nuclear envelope, which isolates the nucleus from the cytoplasm [Katahira et al., 2015].

In the cytoplasm, the stability and translation of mRNAs are regulated by micro RNAs (miRNAs), small RNA molecules that function as guides for another protein complex containing Argonaute family proteins [Meister, 2013,Swarts et al., 2014]. Small interfering RNAs (siRNAs) are exogenous molecules that are exploited in molecular biology for rapid reduction of gene expression, due to their ability to integrate into Argonatute proteins and RNA-induced silencing complexes (RISC), which leads to the specific downregulation of complementary mRNAs [Valencia-Sanchez et al., 2006]. Binding of RISC to mRNAs leads to increased degradation of the mRNA, either by direct cleavage, which requires full sequence complementary and a specific Argonaute protein [Gruber and Zavolan, 2013], or by recruiting deadenylase and decapping complexes that respectively shorten the poly(A) tail and remove the 5' cap of the mRNA, consequently decreasing the mRNA stability [Braun et al., 2012, Gruber and Zavolan, 2013]. RISC complexes loaded with siRNA/miRNA can also silence the translation by preventing the ribosome to assemble on a mRNA [Valencia-Sanchez et al., 2006] but this function has been less well established than the function in mRNA degradation [Gruber

and Zavolan, 2013]. Both siRNA and miRNA are about 21 nucleotides long, but the siRNA are fully complementary to their target whereas miRNAs are more versatile in their binding, with typically only a sequence of 6 to 8 nucleotide, called the 'seed' of the miRNA, being necessary for target recognition. As a result, miRNA can target hundreds of mRNA and most mRNAs can by targeted by miRNAs [Gruber and Zavolan, 2013, Valencia-Sanchez et al., 2006].

Finally, translation is carried out by the ribosome, a large ribonucleoprotein complex composed of two subunits, which bind and assemble the full complex on mRNAs [Ramakrishnan, 2002]. The ribosome carries out mRNA translation, catalysing the formation of the peptide bond between the amino acid carried by an incoming tRNA and the nascent peptide. Translation proceeds one codon at the time via the translocation of the ribosome on the mRNA [Ramakrishnan, 2002]. While the translation machinery has largely been viewed as indiscriminate with regard to the mRNAs, any mRNA having an equal chance of starting translation, this view is currently being challenged [Mauro and Edelman, 2002]. Methods such as ribosome footprinting [REF Ingolia, Weissmann, 2009] revealed that the rate of translation initiation varies widely among mRNAs [Riba et al., 2019]. The rate of translation elongation varies as well, being indirectly affected by the concentrations of tRNAs as well as by the specific codon and amino acid usage of mRNAs [Bulmer, 1991, Vogel et al., 2010, Guimaraes et al., 2014].

1.3 Gene regulatory networks

As the regulators of gene expression (e.g. transcription factors) are themselves the products of gene expression, the relationships between protein-coding gene expression are sometimes summarised as an interaction graph or gene regulatory network (GRN). One of the fundamental questions of molecular biology and of development is how the variety of cell types that define an organism is generated on the basis of the same genetic blueprint. Decades before the constellation of regulators and their mechanisms has been unveiled, the concepts of genetic and epigenetic regulation have been proposed to describe changes in cells that are due to factors encoded in the genome or to factors that do not alter the information contained in the DNA [Waddington, 1942, Tronick and Hunter, 2016]. Conrad Hal Waddington proposed two conceptual analogies to express his understanding of the mechanisms driving cellular differentiation. In his first analogy, epigenetic interactions would create developmental paths similar to a tree structure of railways; cells would evolve on this landscape as a tram moves on a railway network, where switches are analogous to decision points in development, multiplying the number of accessible final stable states [Waddington, 1936]. Later on, the author introduced the idea of an epigenetic landscape [Waddington, 1940], where epigenetic interactions engender an energy-like surface and developmental processes correspond to paths that minimise the energy along their entire course, connecting a naive pluripotent cell to a fully differentiated cell. Again, the developmental process was viewed as similar to marbles rolling down a mountain under the action of gravity towards valleys [Waddington, 1957]. If these 2 analogies have obvious similarities, a major contrast is the deterministic

character of the railways analogy against the relative stochasticity of a marble rolling down valleys. This aspect has broadened the popularity of the later analogy with time and discoveries.

The current view on gene expression is that it occurs in bursts that differ in rate, intensity and length [Yu et al., 2006] among genes and conditions, modulated by specific constellations of TF binding events [Paulsson, 2005]. Combined with the discrete nature of mRNA copy numbers, as well as with the overall low abundance of most mRNA species, the intrinsic noise attributable solely to the stochastic nature of biochemical processes is in the same range as the extrinsic fluctuations caused by regulatory interactions [Elowitz et al., 2002]. TFs are subjected to the same kind of noise in their expression, noise which propagates to the transcriptional targets of these TFs. This leads to heterogeneous responses across a population of cells that are essentially in a similar state [Mcadams and Arkin, 1997], providing a mechanism for generating phenotypic heterogeneity in an isogenic cell population [Elowitz et al., 2002]. Conversely, the remarkable reproducibility of phenotypes emerging from such highly noisy components suggests very strong regulation [Mcadams and Arkin, 1997, Raj and van Oudenaarden, 2008]. This seemingly paradoxical emergence of deterministic macroscopic variables from highly unpredictable microscopic states has been deeply studied in the field of statistical mechanics [Paulsson, 2005]. In fact, the mathematical principles underlying Waddington's epigenetic landscape description of cellular dynamics in development is closely related to the atomic description of statistical mechanics, specifically as it hypothesises the existence of a probability distribution over gene expression states, which underlies differentiation processes as well as the stability of the final differentiated state of cells. Differentiation paths are encoded in the epigenetic landscape as continuous regions of high probability density in the gene expression space. The end point of each such path defines a cell type as a local, stable minimum, where regulatory interactions defined in the neighbourhood of the minima act as stabilisers of the gene expression state against small fluctuations. In light of Waddington's model of an epigenetic landscape, GRNs are powerful objects to explain the mechanisms of cell differentiation, and the astonishing variety of cells in multicellular organisms. Starting from the pluripotent stem cell, internal fluctuations or external inputs affecting the expression of a few specific regulators can influence the expression of other regulators which are themselves influencing downstream regulators. Thus, a cascade of events moves the cell through the gene expression space until the GRN reaches a stable state and a differentiated state is reached.

The concept that noisy gene regulatory interactions generate a macroscopic probability density of accessible regions of gene expression states, borrowed from statistical mechanics, has been further highlighted by published observations that overexpression of a few specific regulators is sufficient to modify the type of a fully differentiated cell [Davis et al., 1987, Kulessa et al., 1995, Xie et al., 2004, Takahashi, 2012], and even to bring it back to a naive, pluripotent stem cell state [Takahashi and Yamanaka, 2006]. These processes, which were called transdifferentiation and reprogramming, respectively, correspond in the Waddington's landscape analogy to the transport of cells above an energy barrier (similar to an activation energy in chemical and nuclear reactions) to a region

of the gene expression space where the landscape will then bring the cell to a stable state, different from the initial state. The mechanisms underlying transdifferentiation and reprogramming are still only partially understood [Takahashi and Yamanaka, 2016]. However, these findings demonstrated that cellular states are plastic. Most importantly, they imply that the information necessary for the differentiation and stabilisation of any cell type is not lost during the process, but is carried in all cells of an isogenic population, at all times, independently of their state. This information presumably resides in the genetically-encoded regulatory networks.

Another aspect of Waddington's rich paradigm of epigenetic regulation is the concept of adaptability, the ability of an organism to change its state to match changes in its environment [Waddington, 1957]. One of the first known and most studied examples of adaptability encoded in a GRN is the *lac* operon in *Escherichia coli* [Jacob and Monod, 1961]. The *Escherichia coli* bacterium needs a specific enzyme for metabolising lactose as a carbon source. The expression of this enzyme is repressed by another protein, which is inactive in the presence of lactose. This simple mechanism allows the bacteria to express the enzyme only when the corresponding metabolite is present. GRNs abound in such feed-forward and feed-backward loops and in more complex sub-network that implement specific dynamic behaviours and provide the modularity necessary for the evolvability of GRNs [Verd et al., 2019]. Indeed, recurrent sub-network designs are used across species, functions and environments [Davidson and Levin, 2005] making possible the development of adaptability in an evolutionary context.

1.4 Experimental quantification of gene expression

Precise knowledge of GRN structure is therefore necessary for understanding the development and functioning of living organisms and it starts with the quantification of gene expression. Different methods have been developed over the years to tackle the challenging task of reliably measuring gene expression.

In the 70's, various methods were developed to monitor the concentration of specific DNA, RNA and proteins products, separating molecular species by size with gel electrophoresis, labelling and visualising the molecules of interest through radioactive or fluorescent markers, or with the help of immunofluorescent antibodies. In Southern blotting, after fragmenting and denaturing DNA and removing the RNA, specific DNA sequences are imaged with a complementary probe that carries a radioactive or fluorescent marker [Southern, 1975]. Conversely, northern blotting visualizes RNAs with similarly labelled complementary probes [Alwine et al., 1977]. Methods to identify specific proteins have also been designed, for example the enzyme-linked immunosorbent assay (ELISA) [Engvall and Perlmann, 1972] and western blotting [Gershoni and Palade, 1983]. Althought these techniques have evolved and been improved over time [Lequin, 2005, Kurien and Scofield, 2006], they are still limited to the quantification of one target DNA, RNA or protein molecule and they are rather imprecise, hardly providing more than a binary description of the expression. A considerable increase in accuracy in the quantification of gene expression came with the reverse transcription polymerase

chain reaction (RT-qPCR), a technique that uses a primer specific to a predefined RNA molecule to amplify this molecule until it is visually detectable, the number of amplification cycles serving as an estimate of expression [Rio, 2014]. This approach also allows multiple genes to be quantified from the same sample, and the degree of gene coexpression to be inferred. Since its early developments in the late 80's [Becker-André and Hahlbrock, 1989], the popularity of RT-qPCR has grown substantially [Freeman et al., 1999]. Nevertheless, the technique still has the drawback that only a relatively small subset of predefined genes can be observed simultaneously, considerably fewer than the thousands of genes expressed in a cell at a given time.

An important breakthrough in the quantification of gene expression came with the development of *Complementary DNA Microarray* technology, which extended the the number of genes whose expression patterns could be measured to virtually genome-wide scale. The approach involved the monitoring of fluorescently labeled cDNA probe hybridisation to cDNAs generated from samples of interest, with gene-specific probes being placed in individual wells [Schena et al., 1995]. As whole genome sequencing enabled the design of probes covering the entire set of genes of a given organism, automatisation of this process eventually allowed the genome-wide quantification of gene expression [Lashkari et al., 1997, Pollack et al., 1999]. This in turn provided valuable insights in the patterns and regulation of gene regulation based on the mRNA level expression [Schulze and Downward, 2001].

About a decade after the dawn of microarrays in mRNA quantification, a number of methods, grouped under the term of next generation sequencing (NGS), have been developed to directly 'read' the DNA or RNA sequences represented in a sample. Due to the substantial scale of several millions and up to billions of small reads that can be observed in a single experiment [Metzker, 2010], as well as to the fact that these methods did not limit identification and quantification to a specific set of molecules, NGS techniques grew rapidly in popularity. They evolved from the broadly used automated Sanger sequencing technique [Metzker, 2010], which used DNA polymerase to elongate a cDNA template with a tuned mixture of deoxyribonucleotide and dideoxynucleotides, the second inhibiting elongation, resulting in synthesised fragments of uniformly distributed lengths that are sorted by size with gel electrophoresis, so that a nucleotide-specific fluorescent dye marking the terminal base in each fragment enables the DNA sequence to be 'read' [Sanger and Coulson, 1975, Sanger et al., 1977, Metzker, 2005]. This method was called *a posteriori* 'first-generation sequencing'. It was for over 40 years and was in fact the method that made it possible to obtain the first draft of the human genome [Collins et al., 2003]. As the human genome was being assembled, drawbacks of this method, particularly with respect to the time and cost, spurred the parallel development of other strategies for sequencing nucleic acids, ultimately leading to what are now called NGS technologies [Metzker, 2005]. They differ in many aspects, including sample preparation, sequencing, imaging and data analysis [Metzker, 2010]. Sample preparation includes fragmentation, PCR amplification and priming of the DNA or RNA material. Sequencing follows, most commonly 'by synthesis', where a DNA polymerase adds fluorescent nucleotide complementary to the template, one by one, allowing their

imaging and identification using lasers [Metzker, 2010]. RNA sequencing (RNA-seq) technologies rapidly took over the field of gene expression estimation, as they clearly improved the sensitivity and the detection of rare transcripts, enabled the distinction between isoforms, increased the dynamic range of detection and the overall scale of the resulting data [Wang et al., 2009, Metzker, 2010].

1.5 Experimental prediction of chromatin state and regulatory elements

Using the methods for gene expression quantification described above, the most common type of analysis is the so-called 'differential gene expression', which aims to identify genes that are expressed at different levels in different samples, as a result of regulatory interactions. Whereas much computational work has been dedicated to the continuous improvement of methods for this analysis [Costa-Silva et al., 2017], differential expression alone does not reveal direct and indirect regulatory interactions. Alternatively, measuring gene expression in time series and calculating covariances provides clues about gene coexpression but such symmetric information alone does not distinguish the regulator from the regulated gene. For a more mechanistic understanding of regulatory interactions that give rise to gene expression patterns, a model-driven approach, that includes other known information connecting genes in a network, is desirable. One such type of information consists of the binding of regulators to the promoters of gene genome-wide, which in turns modulates the rate of transcription.

The genome-wide chromatin opening state can be determined by DNase I hypersensitive site sequencing [Crawford et al., 2006, Boyle et al., 2008] or similarly, by formaldehyde-assisted isolation of regulatory elements followed by deep sequencing (FAIRE–seq) [Giresi et al., 2007]. Both techniques take advantage of the fact that only regions of the chromatin that are accessible to regulator binding are also accessible to and cleaved by DNase I [Elgin, 1988]. Micrococcal nuclease sequencing (MNase-seq), reports the nucleosome location additionally to the open regions using the fact that the regions of the DNA that are covered by nucleosomes are resistant to microccocal nuclease [Cui and Zhao, 2012, Mieczkowski et al., 2016]. The currently most accurate, efficient and popular method for determining chromatin accessibility is the assay for transposase-accessible chromatin using sequencing (ATAC-seq). This technique assesses the positioning of nucleosomes, accessibility of chromatin and TF binding simultaneously [Buenrostro et al., 2013].

To uncover direct interactions between TF and DNA, chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) has been used extensively for more than a decade [Johnson et al., 2007]. In this technique, an antibody specific to a DNA-binding protein is used to extracted genomic regions that are bound by this protein by immunoprecipitation. The bound DNA fragments are then purified and sequenced [Johnson et al., 2007]. Analogous method have been developed to identify binding sites of RNA binding proteins (RBPs) [Wheeler et al., 2018]. Different variants are known such as HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunopre-

cipitation) [Jensen and Darnell, 2008, Darnell, 2010] and PAR-CLIP (photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation) [Hafner et al., 2010]. They use ultraviolet light to crosslink an RBP to RNAs and immunoprecipitation to isolate and purify the RNA sequences to which the RBP is bound. Assayed together with gene expression by RNA-seq, the approaches described here allow to infer regulatory interactions at different levels of regulation (chromatin accessibility, activating or repressing effect of TFs on gene expression, stabilizing or destabilizing effect of RBPs on mRNAs). The complexity and high dimensionality of the system combined with the general noisy and uncertain nature of the experimental data makes an extensive and global inference of GRNs extremely challenging. On the other hand, these experiments do inform about sub-networks of interactions that are active in the studied system.

Several computational approaches have been developed to infer GRNs. They generally use measurements of gene expression to infer regulatory interactions between genes in a reverse engineering approach maximising the patterns and variations explained in the data using different classes of network models. These differ in their deterministic or stochastic nature, in the discrete or continuous nature of assumed regulatory states, in their dynamic or static design [Yaghoobi et al., 2012], [Goutsias and Lee, 2007, Yaghoobi et al., 2012, Chai et al., 2014, Huynh-Thu and Sanguinetti, 2019]. Overall, computational approaches to the analysis of large scale data have provided valuable insights into gene regulatory interactions [Marbach et al., 2012].

The ability of specific proteins to bind DNA and RNA and to regulate the expression of specific genes depends on the strength of physical binding between the protein and the sequence motifs present in the DNA and RNA. This in turn is typically species-, tissue-, and condition-specific, making the complete reliance on experimental measurement (*e.g.* by ChIP-seq and CLIP) impractical [Kundaje et al., 2016, Pan et al., 2019]. Consequently, a large number of methods have been developed to *predict* transcription factor binding sites (TFBS), generally based on experimentally-determined positional weight matrices representing the binding specificity of nucleic acid-interacting proteins [Matys, 2006, Portales-Casamar et al., 2009, Mathelier et al., 2014, Heinz et al., 2010, Kulakovskiy et al., 2013, Weirauch et al., 2014], the sequence context of potential TFBS, the evolutionary conservation of putative binding sites across species, the colocalisation of multiple TFBS or the competition between multiple sites [Bulyk, 2004, Hannenhalli, 2008, Arnold et al., 2012, Jayaram et al., 2016]. Prediction methods are now increasingly benchmarked systematically in projects like the ENCODE-DREAM challenge [Kundaje et al., 2016]. Analogously, computational predictions of functional RBP-binding sites use experimentally validated binding sites to train computational approaches generalising such primary information based on amino acids sequence similarity between proteins, their conservation, their chemical and physical properties, and the protein 3-dimensional secondary structure [Si et al., 2015, Pan et al., 2019]. These methods provide information of various granularity, from simple binary information about the RNA binding property of proteins, to the protein domain involved in the binding, to the 3-dimensional structure of protein-RNA complex and the predicted RNA binding sequence [Zhao et al., 2013].

In contrast to transcription factors or RNA-binding proteins, post-transcriptional

control of gene expression also relies on RNA regulators known as miRNAs [Gruber and Zavolan, 2013, Hausser and Zavolan, 2014]. The rules that govern miRNA-target interactions are distinct from those that govern the interactions of RBPs with RNAs. Consequently, miRNA target prediction uses these principles, relying, among others, on the complementarity of the miRNA 'seed' sequence (nucleotides 2-7 of the miRNA). The conservation of the putative site across species, the binding free energy of the miRNA-mRNA hybridisation computed from the complex structure [Hofacker et al., 1994], and the binding site accessibility defined as the free energy needed to open the structure of the target RNA [Yue et al., 2009, Peterson et al., 2014] are also informative. Different machine learning approaches have been designed to infer miRNA-mRNA interaction rules that are generalisable transcriptome-wide based on experimentally determined targets, which can be determined from gene expression changes that are observed upon transfection of cells with a miRNA [Li and Zhang, 2015, Bradley and Moxon, 2017].

1.6 The single cell resolution in gene expression measurements

The experimental methods described above have all been developed initially to measure gene expression in populations of cells, and to make typical comparisons of populations of cells that differ by the time of their sampling, by the type of tissue being sampled or by some internal (e.g. knock-out of a gene) or external (e.g. change in the environment) perturbation. The differential analysis across time provides insights about the regulators involved in the developmental processes of cell population, global differences across tissue are informative about the tissue specificities, while the changes resulting from a gene knock-out reveal direct and indirect regulatory interaction downstream of the gene of interest.

However, these population-level measurements provide 'average' expression patterns across cells in a population, which may not be representative of any given cell type in the population [Altschuler and Wu, 2010, Wang and Bodovitz, 2010]. Thus, to rigorously tackle the problem of inferring GRNs, measurements at finer resolution are needed. Indeed, returning to the analogy of Waddington, if regulatory interactions are driving the gene expression state of each cell like a landscape constrains the trajectory of a marble, one would need to observe the positions of several individual marbles to learn the shape of the surface. The center of mass of the marbles would not be informative for this purpose. As a matter of fact, even if the assumption a perfectly homogeneous population of cells would hold, the averaging over the population would erase the small fluctuations of cell state around the assumed common stable state that follow from the stochastic nature of gene expression, fluctuation that can reveal regulatory interactions by a simple covariance analysis. Those reasons motivated the development of experimental methods for measuring gene expression in single cells [Eberwine et al., 1992]. Since the first published experiment of RNA sequencing from a single cell a decade ago [Tang et al., 2009], both the number of protocols for RNA-seq at the single-cell resolution (scRNA-seq) and the number of published studies using scRNA-seq never stopped increasing

[Svensson et al., 2018]. The number of cells sequenced in an experiment grew at an impressive rate of roughly one order of magnitude every 1.5 years [Svensson et al., 2018]. Tens of scRNA-seq protocols have been successively developed to allow this exponential growth [Svensson et al., 2017, Ziegenhain et al., 2017], while in parallel, methods were developed to bring the single-cell resolution to ChIP-seq [Gomez et al., 2013, Rotem et al., 2015], ATAC-seq [Buenrostro et al., 2015, Cusanovich et al., 2015], and other sequencing-based approaches to measure numerous different cell state variables [Stuart and Satija, 2019].

Considerable enthusiasm naturally follows the perspectives brought by the wide possibilities of measuring so many aspects of cell state at the single-cell resolution. However, new technologies come with new technical challenges [Stegle et al., 2015]. Whereas a bulk sequencing experiment involves thousands of cell and reports an average state, thereby removing all cell-to-cell variability, only a fraction of this variability is, in fact, informative about the biological state of the cells, as the single-cell level measurements contain substantial technical noise, due to the relatively small quantity of material that can be extracted from a cell [Brennecke et al., 2013]. Sequencing methods rely on the polymerase chain reaction (PCR) [Mullis et al., 1986, Bartlett and Stirling, 2003] to amplify the DNA/RNA signal, which contributes the major part of technical noise [Stegle et al., 2015, Yuan et al., 2017]. In fact, as in a single cell the number of mRNA copies is low for most genes, the PCR predominantly amplifies the sampling noise. This type of noise is much smaller in bulk sequencing, as the material comes from thousands to millions of cells. This complication has been judiciously taken care of by the use of small barcodes, known as unique molecule identifiers (UMI), that are attached to each mRNA molecule before the PCR amplification [Islam et al., 2014], so that the amplified reads coming from the same mRNA molecule can be identified and demultiplexed to obtain the exact number of mRNA molecules that were captured in the each cell. UMIs are nowadays included in the many scRNA-seq protocols for its substantial noise reduction. Nevertheless, the integral number of mRNA molecules thus obtained still carries undesirable bias. First, due to differences in size, themselves coming from both technical (different capture efficiencies in different droplets) as well as biological (differences in the cell cycle stage) factors, the number of mRNAs captured from individual cells in a genetically homogeneous population cells can be quite different [Vallejos et al., 2017]. Such biases are handled by working not with the absolute number but rather with the fraction of each mRNA type per cell, a step commonly called library size normalisation [Stegle et al., 2015, Vallejos et al., 2017]. However, the issue of variability in mRNA copy numbers per cell is still deeper. Even assuming constant and independent (across time and across cells) transcription decay rate for a given gene, at every instant and in every cell, the number of mRNA present in a cell is an integer that follows a Poisson distribution [Grün et al., 2014], due to the stochastic nature of gene expression. Secondly, the fraction of mRNA being captured typically ranges between 10% and 15% [AlJanahi et al., 2018], and up to 30% in the most recent protocols [10X Genomics, 2018]. Hence, assuming a uniform probability for an mRNA molecule to be captured, the capture process corresponds itself to a Poisson sampling, thus adding another component of Poisson noise,

technical in this case [Brennecke et al., 2013, Grün et al., 2014]. The Poisson noise can be dampened by adding a pseudo-count to the normalised count, but this simple method does not rigorously account for the specific nature of the considered noise.

Different strategies have been developed to tackle the problem by fitting a statistical model of gene expression. As the noise on gene expression is at least Poisson, this is the noise that would be measured on a population of cells that have the exact same mean expression level. In a population of cells that does not have the same mean expression level, the variation in gene expression levels beyond Poisson noise are often considered Gamma distributed because of the bursting nature of gene expression [Friedman et al., 2006], ??? I don't understand what this tries to say: a mean expression level that vary across cells, and is often considered Gamma distributed because of the bursting nature of gene expression [Friedman et al., 2006]. ??? however, it is argued that due to the complexity of the reaction networks determining gene expression, not only the noise induced by bursts, but various other sources of multiplicative noise have to be taken into account, and this is better modelled with a log-normal distribution [Beal, 2017]. The Poisson-Gamma mixture has the advantage of being solvable analytically, giving a negative binomial distribution. To infer Gamma-distributed gene expression levels additional to the Poisson noise, different methods are used, like LASSO regression with penalty on non-zero parameters to avoid overfitting [Huang et al., 2018]. It has been argued that scRNA-seq data are *zero-inflated*, in the sense that more zero counts than expected are produced, for biological reasons, as genes not being expressed in a subpolulation of cell, or technical reasons, genes not being captured through sequencing [Vallejos et al., 2017]. Some models take this into account by multiplying the assumed gene expression distribution by a binary probability of a gene to be seen, and infer the model parameters, for instance, by penalised maximum likelihood approach [Risso et al., 2018]. However, it was recently shown that no evidence of zero-inflation is in fact apparent in scRNA-seq data [Svensson, 2020], the high sparsity of the data being simply due to the Poisson nature of the noise, the low expression of many genes, and the low capture rate of the current protocols. Other methods focus primarily on retrieving non-zero expression levels for the *dropout* values [Li and Li, 2018], assuming a Gamma-normal mixture model, where a dropout event is characterise by a Gamma-distribution whereas the other events are characterised by a normal distribution.

Another prevalent approach in analysing single cell data is to assume that these data come from a lower dimensional space referred to as manifold [Lin et al., 2015,Moon et al., 2018]. In fact, the existence of gene regulatory networks implies that only a subspace of gene expression state is accessible, as the expression of every gene depends on the state of its few regulators. Within this scheme, scRNA-seq data can be interpreted as noisy sampling on the manifold of regulator activities, and the task of denoising the data consists in inferring the manifold to which the data can then be projected to 'denoise' it. A popular machine learning technique to infer such manifold is autoencoders. These are artificial neural networks containing several layers of neurons, each layer having a lower dimension than the previous one, down to a *bottleneck* layer, from which a reverse sequence of layers symmetrically brings back the dimension of the first layer

and the original data [Kramer, 1991]. In a machine learning approach, the weights in the neural network are inferred so as to minimise the divergence between the first and the last layers, and it is believed that the bottleneck layer 'captures' the lower dimensional manifold in the desired and predefined dimensionality, so that the final layer brings back a transformation of the original data with reduced noise. Again, models consider zero-inflated Poisson-Gamma mixture distribution but additionally assume that data come from an unknown latent space. The various parameters of the model are learned by inferring the encoder network in a variational Bayesian inference fashion [Lopez et al., 2018], or by stochastic gradient descent using mean square error loss function [Eraslan et al., 2019]. The general problem of dimensionality reduction is evidently much older than scRNA-seq, and a framework tackling this problem, known as diffusion maps [Coifman and Lafon, 2006], is recurrently used in the recent literature to reveal the complex geometrical structure embedded in the gene expression patterns of a large population of cells [Haghverdi et al., 2016]. In this framework, the definition of a distance measure between pair of cells defines a Markov process of cell to cell transition called diffusion operator. The exponentiation of the diffusion operator to the nth power leads to the represention of the transition probabilities between any pair of cells in n steps, producing a multiscale geometrical description of the manifold, with the scale depending on n [Coifman and Lafon, 2006]. The spectral analysis of the diffusion operator provides a representation of the data in a reduced dimension as the diffusion distance is captured by the consecutive eigenvectors, with an accuracy given by the respective eigenvalues [Coifman and Lafon, 2006]. While diffusion maps are mainly used as a visualisation tool and for inferring cellular developmental trajectories, methods have been proposed where this algorithm is used to denoise the data, by projecting the raw data on the inferred manifold [van Dijk et al., 2018]. Other dimensional reduction techniques are used for visualisation, mainly t-SNE [Van Der Maaten and Hinton, 2008], UMAP [McInnes et al., 2018] and PCA [Jolliffe, 2005] (see [van der Maaten et al., 2009, Lin et al., 2015] for comparative reviews on dimensionality reduction). t-SNE maps data points from the gene expression space to a 2 or 3 dimensional space using a gradient descent method, while retaining the local distance structure of the data with a cost function [Van Der Maaten and Hinton, 2008]. UMAP uses theoretical results coming from Riemannian geometry and algebraic topology and guaranteeing that any low dimensional manifold uniformly sampled and represented as simplicial sets can be translated to the topological and underlying metric structure of the manifold, by choosing a metric that approaches the uniformity assumption [McInnes et al., 2018]. PCA is a linear transformation applying a rotation on the data in gene expression space such that each axis of the new base is the linear combination of the genes that recursively maximises the variance of the data along the axis and is orthogonal to all previous axes [Jolliffe, 2005].

1.7 Subjects covered in this thesis

During this PhD, I worked on three different aspects in the broad field of experimental and theoretical analysis of gene regulation.

The first part, *Quantifying the strength of miRNA-target interactions*, addresses the problem of predicting mRNA targets of miRNAs. I show that biochemical measurements of miRNA-mRNA interactions can be used to optimise the parameter inference of a pre-existing model of miRNA target prediction. This model named MIRZA [Khorshid et al., 2013], predicts miRNA-mRNA binding using 25 energy parameters that describe the miRNA-mRNA hybrid structure, with 2 base pairing parameters for the AU and GC pairs, 3 configuration parameters for the symmetric and asymmetric loops, and 21 positional parameters for the 21 nucleotides of the miRNA sequence. MIRZA was built to infer these parameters from Argonaute protein CLIP data, which captures potential targets of miRNAs. Upon the publication of precise measurements of chemical kinetic constants of miRNA-mRNA binding interactions between a mRNA target and a set of systematically mutated miRNA sequences [Wee et al., 2012], we reasoned that such data could be used to improve the parameters inference of the MIRZA model. After showing that the prediction of the existing model on the set of measured miRNA-mRNA pairs shows high correlation with the binding energy calculated from the measurements, I used simulations as a proof of principle of the inference procedure and to design measurements that would be needed to infer the parameters of the MIRZA model.

Staying in the field of miRNA, in *Single cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction*, I developed an approach to infer miRNA targets based on scRNA-seq data from cells that express the miRNA at different levels. A miRNA can target several hundreds of different mRNAs and is present in the cell in limited quantities, implying that the interaction of a target mRNA with a specific miRNA depends on its concentration and on the interactions of the miRNA with its other targets. In other words, since miRNA binding is exclusive, mRNA targets compete for the same miRNA pool. Therefore, the concentrations of the thereby coupled mRNAs depend not only on the miRNA concentration but also on the concentration of every competing mRNA that is targeted by the same miRNA. To study this, HEK 293 cell lines were constructed to inducibly express a miRNA (hsa-miR-199a) as well as the mRNA encoding a green fluorescent protein. Express from the same promoter as the miRNA, this mRNA allows the monitoring of the miRNA concentration. The study aimed not only to determine the parameters of individual mRNA-mRNA interactions, but also to assess the dgree to which mRNAs act in a competitive manner to influence each other's expression. scRNA-seq was chosen to bring the resolution needed to reach these goals. The effect of the miRNA on a bound target is to increase its decay rate, hence the expression levels of the targets depends on the miRNA concentration and their binding energy. To gain insight into the target binding energy, we constructed a model considering mRNA transcription rate, the miRNA-mRNA binding/unbinding rate, the mRNA decay rates in the bound and unbound state, and the free/bound concentration of miRNA. We showed that the model can be factored in terms of the miRNA concen-

trations in individual cells and the miRNA-mRNA target interaction parameters and we solved the model to obtain estimates of miRNA-mRNA interaction parameters, which we showed explain the mRNA levels in cells more accurately than the sequence-based computationally predicted interaction energies.

Finally, in *Bayesian inference of gene expression levels in single cells* I carried out fundamental technical work on the normalisation of count data obtained in scRNA-seq experiments. As introduced above, multiple strategies have been developed with the aim of reducing the high level of noise present on such data, and estimating a 'true' biological state of expression for each gene in each cell. While the project aimed to reconstruct the Waddington landscape of regulator activity based on the single cell gene expression measurements, at the start of the project we realised that there is no satisfactory solution to gene expression normalisation in single cells in the literature. Thus, we tackled this problem with a Bayesian model, considering each gene independently and inferring a posterior probability of gene expression in each cell. Our model assumes a log-normal distribution of gene expression across cells and additional Poisson noise caused by the stochastic process of gene expression and the sampling process introduced by the mRNA capture in experimental protocols. These normalised gene expression values are the basis of a motif-activity response based approach for inferring the activity of TFs and miRNAs in individual cells, and for reconstructing the underlying landscape.

Chapter 2

Quantifying the strength of miRNA-target interactions

Jeremie Breda, Andrzej J. Rzepiela, Rafal Gumienny,
Erik van Nimwegen and Mihaela Zavolan

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50–70, 4056 Basel, Switzerland

Methods 85 (2015) 90–99

Abstract:

We quantify the strength of miRNA-target interactions with MIRZA, a recently introduced biophysical model. We show that computationally predicted energies of interaction correlate strongly with the energies of interaction estimated from biochemical measurements of Michaelis-Menten constants. We further show that the accuracy of the MIRZA model can be improved taking into account recently emerged experimental data types. In particular, we use chimeric miRNA-mRNA sequences to infer a MIRZA-CHIMERA model and we provide a framework for inferring a similar model from measurements of rate constants of miRNA-mRNA interaction in the context of Argonaute proteins. Finally, based on a simple model of miRNA-based regulation, we discuss the importance of interaction energy and its variability between targets for the modulation of miRNA target expression *in vivo*.

Highlights:

- We describe a framework for inferring parameters of a biophysical model of miRNA-target interaction from a variety of medium and high-throughput data sets.
- Of the variant models that we inferred in this work, the most effective in predicting functional miRNA targets is the MIRZA-CHIMERA model, which is derived from chimeric miRNA-mRNA sequences that were captured in Argonaute crosslinking and immunoprecipitation (CLIP) experiments.
- While the captured chimeric sequences suggested that several miRNAs target predominantly non-canonical sites, the MIRZA model inferred from the chimeras of these miRNAs does not improve the genome-wide prediction of their targets.

Keywords:

miRNA, MIRZA, CLIP, CLASH, non-canonical miRNA binding, miRNA target prediction

2.1 Introduction

MicroRNAs (miRNAs) have emerged as important regulators of gene expression across a wide range of species. They are endogenously encoded small RNAs that are incorporated in ribonucleoprotein complexes also containing an Argonaute (Ago) protein, which they guide to other RNA targets to modulate their expression [Huntzinger and Izaurralde, 2011]. Although comparative genomic analyses

indicate that a miRNA has on average hundreds of targets [Lewis et al., 2005], how these predicted targets respond to changes in miRNA concentration is not entirely clear. The best-documented outcome of miRNA-target interaction is target destabilization [Eichhorn et al., 2014], which is typically modest, but can give rise to interesting behaviors of miRNA-containing regulatory networks. These include the 'threshold-linear' response of miRNA targets to their transcriptional induction [Levine et al., 2007, Mukherji et al., 2011] and the ultrasensitivity of target expression to the miRNA concentration [Buchler and Louis, 2008]. The steady-state level of a given mRNA reflects the balance between transcription and decay. If the mRNA decay rate were constant, not modulated by miRNAs, the mRNA level would be expected to increase linearly with the transcription rate. However, if transcriptional induction occurs in the presence of a cognate miRNA, the target is expected to respond in a 'threshold-linear' manner: when the transcription rate is low, the few mRNA molecules that are produced are bound by the cognate miRNA and degraded. Once the transcription rate is sufficiently high for the mRNAs to saturate the miRNA-Ago complexes, the mRNAs escape the miRNA-induced repression and accumulate at a rate proportional to their transcription rate. The location of the threshold depends on the abundance of miRNA-Ago complexes, while the steepness of the transition between the two regimes depends additionally on the affinity of miRNA-target interaction.

We can illustrate these concepts with a simple model that focuses on the interaction of a single miRNA target with the miRNA and on the effect of this interaction on the rate of target decay, ignoring the possible effect of miRNAs on translation, the possible competition between targets for miRNAs and vice versa, other secondary effects such as feedbacks on target transcription rates, etc. Although these aspects most likely are relevant in *in vivo* situations, they go beyond the scope of our present study. Let us consider a miRNA target that is transcribed at rate $\alpha[mol \cdot s^{-1}]$ and decays with rate $\delta[s^{-1}]$. The free miRNA target $F[mol]$ associates at rate $\beta[mol^{-1} \cdot s^{-1}]$ with miRNA-Ago complexes whose total concentration in a cell we assume to be constant, Σ. This leads to the formation of ternary target-miRNA-Ago complexes whose concentration we denote by $A[mol]$, which can either dissociate into their components with rate $Q[s^{-1}]$, or fall apart due to the degradation of the miRNA target, which occurs at rate $d\delta[s^{-1}]$. The dynamics of these molecular species can then be described by the following equations:

$$\frac{dF}{dt} = \alpha - \delta F - \beta(\Sigma - A)F + \rho A \tag{2.1}$$

$$\frac{dA}{dt} = \beta(\Sigma - A)F - \rho A - d\delta A \tag{2.2}$$

Solving this system of differential equations we obtain the dependency between the concentration of the free (and total) target and its transcription rate, which

has the threshold-linear form. Fig. 2.1 shows how the concentration of the free mRNA target responds to changes in target transcription rate, assuming values for the parameters $\delta = 0.11/hour$ and $d = 1.55$, which we have recently estimated [Hausser and Zavolan, 2014]. To illustrate the expected behavior of high and low affinity targets we use two distinct values of the rate of ternary complex formation β, namely 0.24 and 2.4 cell/molecule/hour, and two distinct values of the rate of ternary complex dissociation ρ, namely 2.16 and 21.6 1/hour. To further explore the behavior of targets of low, intermediate and high abundance miRNAs, we consider three total concentrations Σ of miRNA-Ago complexes, namely 10, 100 and 1000 molecules/cell. Our model thus assumes that the total concentration of miRNA-Ago complexes (free or bound to targets) is constant and does not respond to changes in miRNA target concentration. Although it remains unclear whether this assumption holds *in vivo*, data showing that the targets of endogenous miRNAs are up-regulated in response to transfection of exogenous siRNAs [Khan et al., 2009] suggest that at least the number of Argonaute molecules in a cell does not scale with the number of small RNAs that are present in cells. It can be observed that the transcription rate at which the target escapes miRNA regulation and accumulates rapidly depends on the total concentration of miRNA-Ago complexes, and that the transition is sharper for targets that have a higher rate of association with miRNA-Ago complexes. These behaviors have been observed in experiments with reporter constructs [Mukherji et al., 2011, Bosson et al., 2014].

So far we discussed the expected behavior of an individual miRNA target. However, because a miRNA probably has hundreds of targets, one of the strongly debated questions in the field is whether changes in expression of one of these targets affects the expression of the others by modulating their interaction with the common targeting miRNA. Computational studies have shown that the targets of a miRNA are expected to respond in an asymmetrical manner, changes in expression of high-affinity targets affecting the binding of the lower affinity targets but not the other way around [Figliuzzi et al., 2013, Bosia et al., 2013]. Whether these behaviors indeed occur *in vivo* is largely unknown. Rather, it has become clear that progress in understanding the impact of miRNAs on gene expression requires accurate measurements of miRNA abundance in single cells, estimates of the number of binding sites that a miRNA typically accesses within a cell, and estimates of the affinity of interaction between a miRNA and its multiple targets.

The abundance of individual miRNAs in mammalian cells varies over orders of magnitude (see for e.g. [Bissels et al., 2009]). MiR-122, a highly-expressed, hepatocyte-specific miRNA can reach 66'000 copies per cell in mouse liver cells and 135'000 in primary human hepatocytes [Chang et al., 2004]. The more typical range for well-expressed miRNAs is 1'000-10'000 molecules per cell [Bissels et al., 2009], which can probably be accommodated by the population of Ago proteins, whose abundance per cell has been estimated to be $\sim$ 140'000-170'000 molecules (in a mouse epidermis and a human melanoma cell) [Wang et al., 2012].

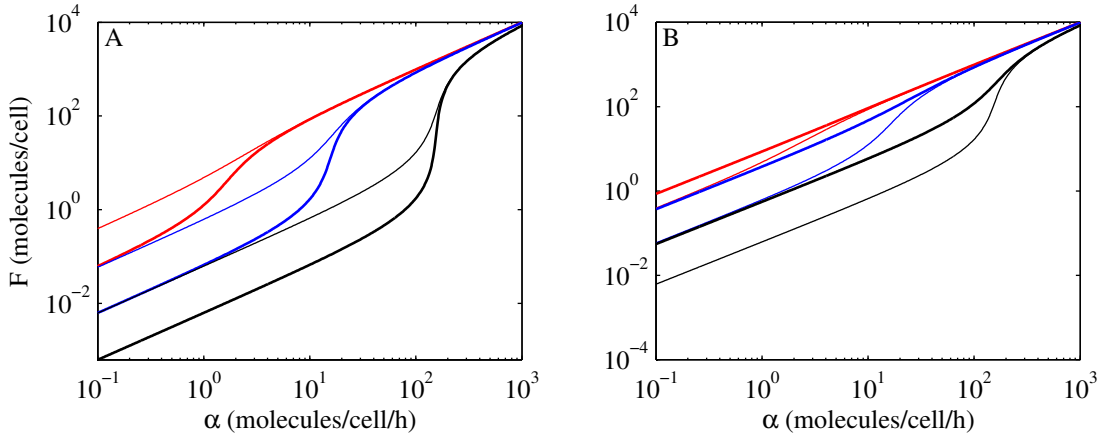


Figure 2.1: Accumulation of miRNA targets as a result of increasing transcription, in the presence of miRNAs, based on the steady state solution of equations (2.1) and (2.2). The three colors correspond to three total concentrations of miRNA-Ago complexes of 10 (red), 100 (blue) and 1000 (black) molecules/cell. (A) Thin lines correspond to low rates of target-miRNA-Ago association $\beta = 0.24$ cell/molecules/hour, and thick lines to 10-fold higher association rates, $\beta = 2.4$ cell/molecules/hour, with $\rho = 2.16$ 1/hour. (B) Thin lines correspond to low rates of target-miRNA-Ago dissociation of $\rho = 2.16$ 1/hour, and thick lines to 10-fold higher dissociation rates, $\rho = 21.6$ 1/hour, with $\beta = 0.24$ cell/molecules/hour.

The number of target sites that a miRNA can access within an individual cell remains hotly debated [Bosson et al., 2014]. Recently developed methods have enabled quantification of mRNA species within single cells, although the mRNA capture rate appears to be low, around 10% [Grün et al., 2014]. A cursory analysis of the published mouse embryonic stem cell (ESC) single cell data shows that among the mRNAs that were captured, miRNA targets occur in a handful of copies such that the top 100 predicted targets of individual miRNAs yield a few hundred captured target molecules per cell (Fig 2.2). The targets of the mouse ESC-specific miR-294 are less abundant, ~ 1 captured mRNA per cell, compared to targets of the ubiquitously expressed miR-16 and of some miRNAs that are expressed in differentiated tissues (e.g. the general differentiation marker let-7, the neuron-specific miR-124, the muscle-specific miR-1 and the epithelia-specific miR-200a), which were captured in 2-5 copies, on average. Assuming a capture rate of 10%, a mouse ESC thus expresses on average 10-50 molecules per miRNA target. The argument can be made that our estimation ignores the fact that ESCs already contain miRNAs which have reduced the levels of their targets and that we have thus underestimated the number of miRNA targets. Indeed, to improve these estimates we would need to quantify mRNA abundance in ESCs devoid of miRNAs (Drosha/Dicer knock-out ESCs). However, many studies in which miRNAs have

been transfected in cells in which they were not previously expression found only modest changes (less than 2-fold) in target levels and thereby decay rates (see for e.g. [Hausser and Zavolan, 2014]). If a miRNA does target over a hundred distinct mRNA species, binding to perhaps multiple sites within a mRNA, the number of putative binding sites of a miRNA in a single cell can reach $10^3 - 10^4$. Precise estimates of the number of binding sites and the ratio of binding sites to miRNA-Ago molecules are essential for understanding the behavior of the targets *in vivo*, in individual cells.

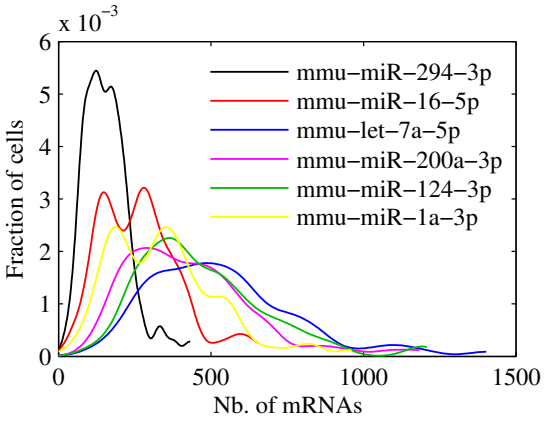


Figure 2.2: Distribution of the number of targets of individual miRNAs that were captured from individual ESCs [Grün et al., 2014]. For each miRNA, the number of molecules of top 100 targets that were predicted with the seed-MIRZA-G-C miRNA target prediction program [Gumienny and Zavolan, 2015] were counted. The actual number of molecules was probably 10-fold higher (assuming that the capture rate of mRNA molecules in mRNA-seq is ∼10%).

2.2 Inferring the strength of miRNA-target interactions from experimentally-determined target sites; theory

An important breakthrough in the experimental identification of miRNA targets came with the development of methods based on the crosslinking and immunoprecipitation of Argonaute proteins (Ago-CLIP) [Chi et al., 2009, Hafner et al., 2010], which enabled the capture of *in vivo* miRNA targets in high-throughput. The basic principle is to crosslink proteins to RNAs in vivo with ultraviolet light, immunoprecipitate the protein of interest and associated RNAs with a specific antibody, and prepare the protein-bound RNA fragments for deep sequencing. The resulting reads can be used not only to identify the mRNAs that were bound by miRNA-guided Argonaute proteins, but also to learn more about how miRNAs interact with their targets. For example, to describe this interaction, in previ-

ous work we introduced a model (MIRZA) that includes besides parameters for $A-U$, $G-C$, and $G-U$ base pairs, for symmetrical and asymmetrical loops, a set of parameters corresponding to miRNA position-dependent contributions to the interaction energy [Khorshid et al., 2013]. The latter could result from the interaction taking place within the context of the Argonaute protein (Fig. 2.3). Parameter values were inferred within a probabilistic framework, by maximizing the likelihood of the CLIP data. They confirmed the known importance of the miRNA 5' end (also known as 'seed' [Lewis et al., 2005]) in the interaction with the target. However, application of the model to the CLIP sites suggested that many are bound in a 'non-canonical' manner (i.e. without perfect complementarity to the miRNA seed) and that the proportion of non-canonical sites that were captured for a given miRNA with CLIP increased with the abundance of the miRNA [Khorshid et al., 2013]. Because MIRZA provides a quantitative measure of the strength of interaction of miRNAs with target sites, it can be used not only for genome-wide prediction of binding sites but also to study miRNA-dependent regulation in deeper quantitative detail. In a parallel development, a next step in the experimental identification of miRNA target sites has been taken with the simultaneous capture of interacting miRNAs and target sites as chimeric sequence reads [Helwak et al., 2013, Grosswendt et al., 2014]. Initial analysis of these data suggested that miRNAs may differ in their mode of interaction with the targets.

Thus, important open questions for the quantitative modeling of miRNA-target interactions are: what approach yields the most predictive model; what structure does this model have; are miRNA-specific models necessary to explain the experimental data? In the following we describe the miRNA-target interaction models that we inferred with the MIRZA approach from various types of high-throughput data, and we evaluate their ability to identify functional miRNA targets, that are destabilized upon transfection of the cognate miRNA.

2.2.1 Input data: Argonaute-bound RNA fragments. Output: General model of miRNA-target interaction MIRZA-CLIP

A target site m of a miRNA μ can be in one of two states, namely bound or unbound to the miRNA. Denoting the energies of the bound and unbound states by E_B and $E_{\bar{B}}$, the probability to find the site in bound state will be given by $P_B = \frac{e^{E_B}}{e^{E_B}+e^{E_{\bar{B}}}}$. The 'bound' state consists in fact of all ways in which the miRNA is hybridized with the target in the context of the Ago protein. Denoting by $E(m,\mu,\sigma)$ the energy of the state in which site m is bound to miRNA μ in configuration σ, e^{E_B} is proportional to $\sum_\sigma e^{E(m,\mu,\sigma)}$. Similar to the standard model of RNA-RNA interaction [Xia et al., 1998], $E(m,\mu,\sigma)$ can be written in terms of a small number of parameters such as the energy of $A-U$, $G-C$ and $G-U$ base pairs, the energy for opening a loop in the miRNA-target hybrid, energies for extending a loop by a nucleotide in the miRNA, or in the mRNA, or by two un-

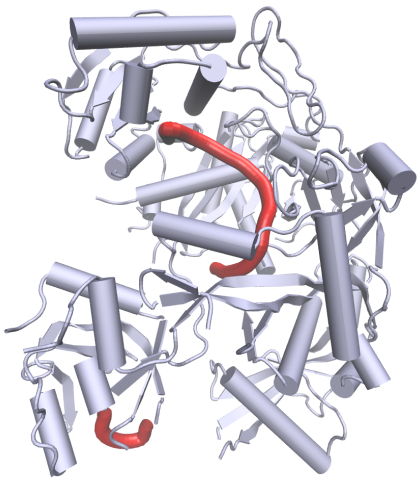


Figure 2.3: Crystal structure of the human AGO-2 protein (silver) in complex with miR-20a (red) [Elkayam et al., 2012]. The 'seed' nucleotides are visible in the structure because the conformational entropy of the miRNA 5' end in the binding pocket of AGO-2 is limited. The residues 11-16 of the miRNA are not resolved due to their conformational freedom. The terminal 3' end nucleotides, that contribute to the anchoring of the miRNA within AGO-2, are again visible.

paired nucleotides in the miRNA and target. In addition, specific to the MIRZA model of miRNA-target interaction [Khorshid et al., 2013] is a set of miRNA-position-specific energies (Fig. 2.4). The logarithm of the 'quality score' of a site for a miRNA that MIRZA computes can be viewed as the energy of interaction between the miRNA and the target. An efficient dynamic programming algorithm for computing target quality scores has been proposed [Khorshid et al., 2013]. This enables one to infer the parameters of the MIRZA model by maximizing the likelihood of the Ago-CLIP data. Here we have repeated the analysis of the $\sim$ 3000 Ago2-CLIP sites that were reproducibly isolated in multiple CLIP experiments [Khorshid et al., 2013, Kishore et al., 2011] to derived the baseline MIRZA-CLIP model shown in Figure 2.4.

2.2.2 Input data: chimeric miRNA-mRNA sequence reads. Output: General model of miRNA-target interaction MIRZA-CHIMERA

As mentioned in the Introduction, Helwak et al. [Helwak et al., 2013] designed the Crosslinking and Sequencing of Hybrids approach (CLASH), in which the interacting RNAs are ligated prior to sequencing, thereby enabling the simultaneous capture of interacting miRNAs and target sites. These appear as "chimeric reads"

MIRZA-Class I	let-7a-5p, let-7e-5p, let-7f-5p, miR-10a-5p, miR-10b-5p, miR-125a-5p, miR-125b-5p, miR-1260b, miR-1301-3p, miR-130b-3p, miR-15b-5p, miR-17-5p, miR-183-5p, miR-185-5p, miR-23a-3p, miR-27b-3p, miR-31-5p, miR-324-3p, miR-339-5p, miR-34a-5p, miR-423-5p, miR-455-3p, miR-484, miR-744-5p
MIRZA-Class IV	miR-181b-5p, miR-221-3p, miR-30c-5p, miR-30d-5p, miR-320a, miR-361-5p, miR-92a-3p, miR-92b-3p

Table 2.1: Chimeras of the indicated miRNAs, obtained from the data set of Grosswendt et al. [Grosswendt et al., 2014] were used to infer MIRZA-Class I and MIRZA-Class IV models.

each composed partly of a miRNA and partly of the miRNA target. Grosswendt et al. [Grosswendt et al., 2014] subsequently reported that a substantial number of ligated miRNA-target site chimeras can be found even in samples prepared with a standard CLIP protocol. In contrast to Ago-CLIP, in these data sets there is no uncertainty about the miRNA that guided the interaction with each target site captured in the chimeras. Thus, in maximizing the likelihood of the data to infer a MIRZA-type model, one only needs to sum over all the ways in which the miRNA and target site in each chimera hybridizes with each other (and not over the miRNAs that could have interacted with the target site, as in the case of Ago-CLIP sites). We used the miRNA-target site pairs that were inferred by Grosswendt et al. from various PAR-CLIP and HITS-CLIP experiments (Table 2.1 and Supplementary Table 3 in [Grosswendt et al., 2014]) to construct a general model that could explain all these interactions. We called this model MIRZA-CHIMERA. Compared to the MIRZA-CLIP model that we inferred from Ago-CLIP data , MIRZA-CHIMERA seems to put less emphasis on the miRNA seed (Fig. 2.4). The functional relevance of these differences will be discussed in the following sections.

2.2.3 Input data: chimera of a specific miRNA with target sites. Output: miRNA-specific model of interaction with the target

The CLASH study reported that some miRNAs, such as miR-92a and miR-181b, interact with their targets predominantly through their 3' rather than the 5' end, yielding 'class IV' chimeras [Helwak et al., 2013]. Other miRNAs such as those of the let-7 family were captured rather in 'class I' chimeras, in which the miRNA presumably interacted through the 'seed'. These observations suggest that the accuracy of miRNA target prediction could be improved through the use of miRNA-specific models of interaction. We decided to test this hypothesis here. However, because the available data sets [Helwak et al., 2013,Grosswendt et al., 2014] contain a limited number of distinct target sites ligated to individual miRNAs, we inferred

'class'-specific rather than miRNA-specific models. Concretely, from the data of Grosswendt et al. [Grosswendt et al., 2014] we selected a total 2589 chimeras of 24 miRNAs (those that yielded predominantly class I chimeras in the data of Helwak et al. [Helwak et al., 2013]) to train the "MIRZA-Class I" model and 949 chimeras of 8 miRNAs (those that yielded predominantly class IV chimeras) to train the "MIRZA-Class IV" model. The corresponding miRNAs are listed in Table 2.1. The parameters of these models, shown in Figure 2.4, indicate a positive contribution of the seed positional parameters in the MIRZA-Class I model, but not in the MIRZA-Class IV model. However, Figure 2.4 also shows a trend of positional parameters to progressively decrease from the seed to the 3' end in the MIRZA-Class IV model, but not in the MIRZA-Class I model. We test the functional relevance of these differences in a subsequent section.

It has been recently observed that the miRNAs that were reported to form class IV hybrids have G/C-rich 3' ends [Wang, 2014]. We reproduced these observations here (Fig. 2.5). Furthermore, we found that the proportion of class I hybrids that were captured for a miRNA decreases with the G/C content of the miRNA 3' end, while the proportion of class IV hybrids shows the opposite trend (Fig. 2.5). A possible explanation behind the different propensities of different miRNAs to yield class I or class IV chimeras is that the G/C-content of the miRNA 3' end stabilizes the interaction with the target site, facilitates ligation and leads to an over-representation of this type of sites among the chimeric sequences. This possibility would need to be investigated in more detail before miRNA-specific modes of interaction are inferred from chimera data.

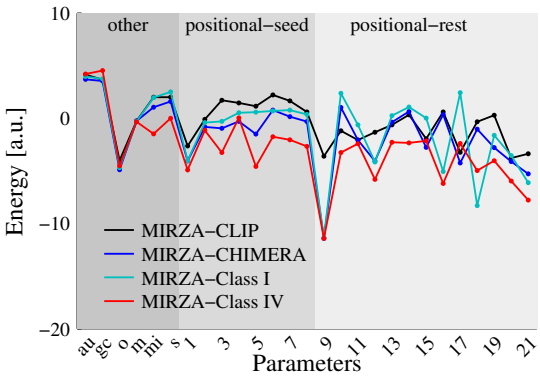


Figure 2.4: The 27 parameters of various MIRZA model variants. From left to right, base-pair parameters ($A-U, G-C, G-U = 0$), loop parameters (o: opening a loop, m: looped out mRNA nucleotide, mi: looped out miRNA nucleotide, s: symmetrical loop) and the 21 positional parameters are shown. The parameters of the MIRZA-CLIP model are shown in black, those of the MIRZA-CHIMERA model in blue, those of the MIRZA-Class I model in cyan and those of the MIRZA-Class IV model in red.

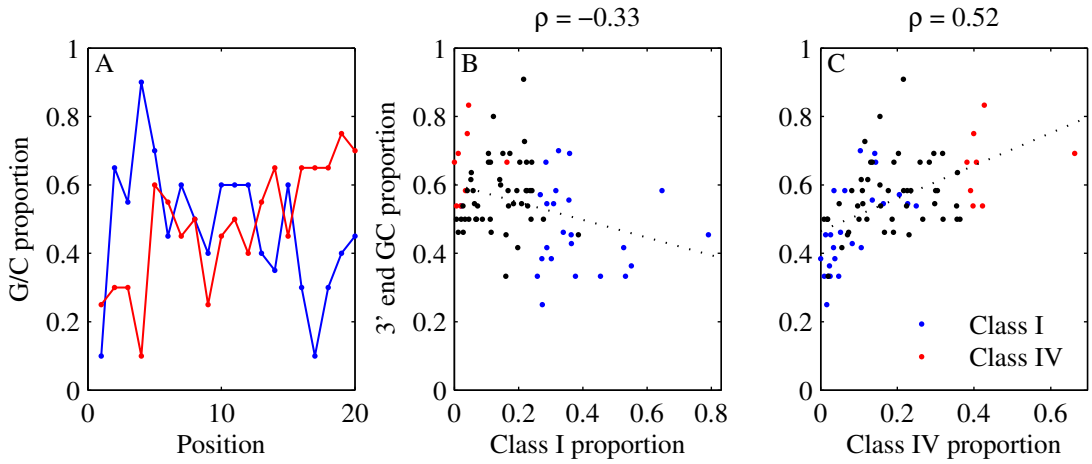


Figure 2.5: Relationship between the nucleotide composition of the miRNA and the type of hybrids in which the miRNA was captured. The miRNAs used to infer the MIRZA-Class I model are shown in blue, the miRNAs used to infer the MIRZA-Class IV model are shown in red and other miRNAs are shown in black. Data for analysis taken from Helwak et al. [Helwak et al., 2013]. (A) Proportion of G/C nucleotides at different positions along miRNAs that yield predominantly class I and IV hybrids/chimeric reads in the data set of Helwak et al. [Helwak et al., 2013]. (B) Correlation between the proportion of G/C nucleotides at the 3' end of a miRNA and the proportion of captured class I chimeras. (C) Correlation between the proportion of G/C nucleotides at the 3' end of a miRNA and the proportion of detected class IV chimeras.

2.3 Results

2.3.1 Evaluating the models on biochemical data

The 'quality score' assigned to a site by the MIRZA model takes into account all possible configurations in which the miRNA can hybridize to the target site within the ternary miRNA-target site-Ago complex, and provides an estimate of the binding energy between the miRNA and the target site. Thus, if the model is accurate, it should be able to predict the free energy of interaction determined with biochemical approaches. The dissociation constant K_D, which is the ratio of the rates of dissociation (k_{off}) and association (k_{on}) of molecules in a complex, $K_D = \frac{k_{off}}{k_{on}}$, should be related to the Gibbs free energy of interaction through the relationship $\Delta G = -k_B T \log\left(\frac{1}{K_D}\right)$, where k_B is the Boltzmann constant and T is the temperature. Although only few measurements of miRNA-target dissociation constants are available, particularly for mammalian systems, Wee et al. [Wee et al., 2012] measured a related constant, namely the Michaelis-Menten constant. This is defined as $K_M = \frac{k_{cat}+k_{off}}{k_{on}}$, thus including besides the dissociation and association

rates the rate with which the miRNA catalyzes the target cleavage. Wee et al. measured for K_M's for perfectly complementary sequences (PM) and for sequences that have mismatches at different positions along the miRNA (MM) in the context of Argonaute 1 protein of Drosophila melanogaster [Wee et al., 2012] and then correlated $\log\left(\frac{K_M^{PM}}{K_M^{MM}}\right)$ with the difference in the free energy of interaction of the the perfectly matched and mismatched hybrids given by the RNAStructure software [Reuter and Mathews, 2010]. Computing this correlation separately for duplexes in which mismatches were located at the 5' and 3' ends of the miRNA, respectively, Wee et al. concluded that the standard base pairing rules apply to miRNA-Ago2-target complexes [Wee et al., 2012]. We thus sought to use the measurements of Wee et al. [Wee et al., 2012] to further validate the MIRZA models that we inferred from CLIP data sets.

First, we compared the energy differences inferred from measurements of K_M's with those predicted with the current version (5.7) of the RNAStructure software and with those predicted with MIRZA-type models. As described by Wee et al. [Wee et al., 2012], we found relatively good correlations between RNAStructure-based predictions and experimental measurements, if we consider separately hybrids with mismatches in the miRNA seed region (Spearman correlation coefficient $\rho = 0.81$, p-value $= 0.015$) and in the miRNA 3' end (Spearman correlation coefficient $\rho = 0.4$, p-value $= 0.20$). However, considering all the hybrids together, the correlation is rather poor (Spearman correlation coefficient $\rho = 0.2$), presumably because the nearest neighbor model implemented in RNAStructure does not appropriately describe interactions that take place within RNA-protein complexes, where different nucleotides in the RNA can have disproportionate contributions to the energy of interaction.

In contrast, evaluating all of the hybrids within the MIRZA-CLIP model yields predictions that are strongly correlated with the experimental results (Spearman correlation coefficient $\rho = 0.85$, p-value $= 3.6e-9$, 95% confidence interval $= [0.71, 0.93]$). Interestingly, the MIRZA-CHIMERA model gives a slightly higher correlation with the experimental data (Spearman correlation coefficient $\rho = 0.87$, p-value = 3e-9, 95% confidence interval $= [0.73, 0.94]$), although the difference is not significant. Thus, these two models, that were inferred from different types of sequenced miRNA target sites, predict remarkably well the energies of interaction between miRNAs and target sites that are inferred from biochemical measurements (Fig. 2.6).

2.3.2 Genome-wide prediction of miRNA targets

One of the main applications of these models is in the genome-wide prediction of miRNA binding sites. However, the predicted energy of interaction between a miRNA and a target site is only one of the factors that contributes to a functional interaction. Other features of the target site have also been shown to be

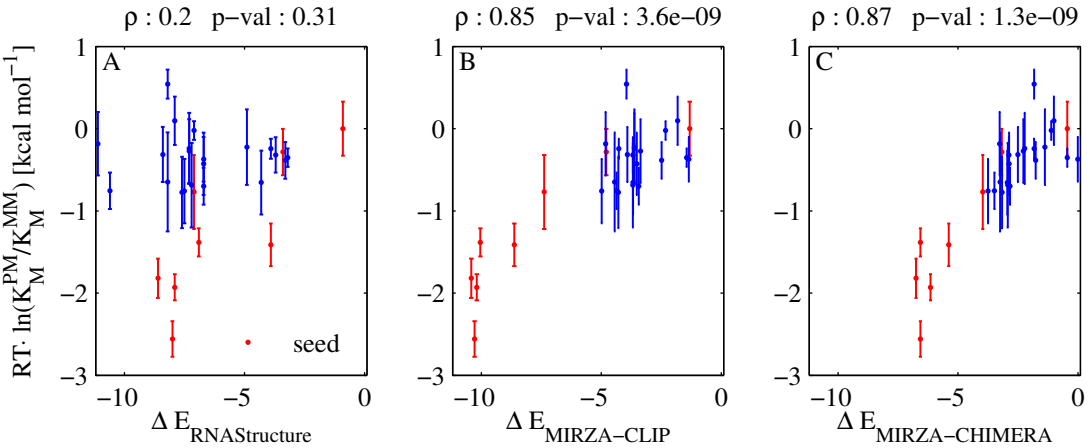


Figure 2.6: Ratio of binding free energies of mismatched and perfectly matched hybrids. The Spearman correlation was computed between the values estimated based on biochemical measurements (energy of interaction $\ln\left(\frac{1}{K_M}\right)$) and values predicted with three distinct models: RNAStructure 5.7 (left), MIRZA-CLIP (center) and MIRZA-CHIMERA (right). Data points in red correspond to hybrids with mismatches in the miRNA seed region, those in blue to hybrids with mismatches in the 3' region.

important [Hausser et al., 2009]. Thus, in recent work we sought to build on MIRZA and develop a model that is suitable for accurate prediction of miRNA binding sites genome-wide. The resulting MIRZA-G model combines the MIRZA target quality score with the accessibility of the target site, the G/U content of the region in which the site is embedded, the relative location of the site in the transcript and, optionally, with the degree of evolutionary conservation of the putative target site (Fig. 2.7). MIRZA-G is trained by fitting a generalized linear model with a logit function to discriminate between miRNA-complementary sites located in mRNAs that do and mRNAs that do not respond to the transfection of the cognate miRNAs [Gumienny and Zavolan, 2015]. Furthermore, because high-throughput studies evaluate the effects of miRNAs at the level of transcripts and genes rather than individual sites, MIRZA-G computes transcript/gene scores, summing up the probabilities that individual target sites have a functional impact. Using different MIRZA variants to compute target quality scores for the MIRZA-G model we can test the ability of these variants to predict which transcripts are most affected by the transfection of individual miRNAs. Thus, we employed the MIRZA-CLIP/CHIMERA/Class I/Class IV models individually within the MIRZA-G framework to predict and rank targets of individual miRNAs. Because different MIRZA variants yield different distributions of target quality scores and in the genome-wide prediction of target sites we only consider putative sites with a minimal 'target quality' score, we have used different thresholds for different mod-

els. The weight of different features of target sites within the MIRZA-G model were kept unchanged.

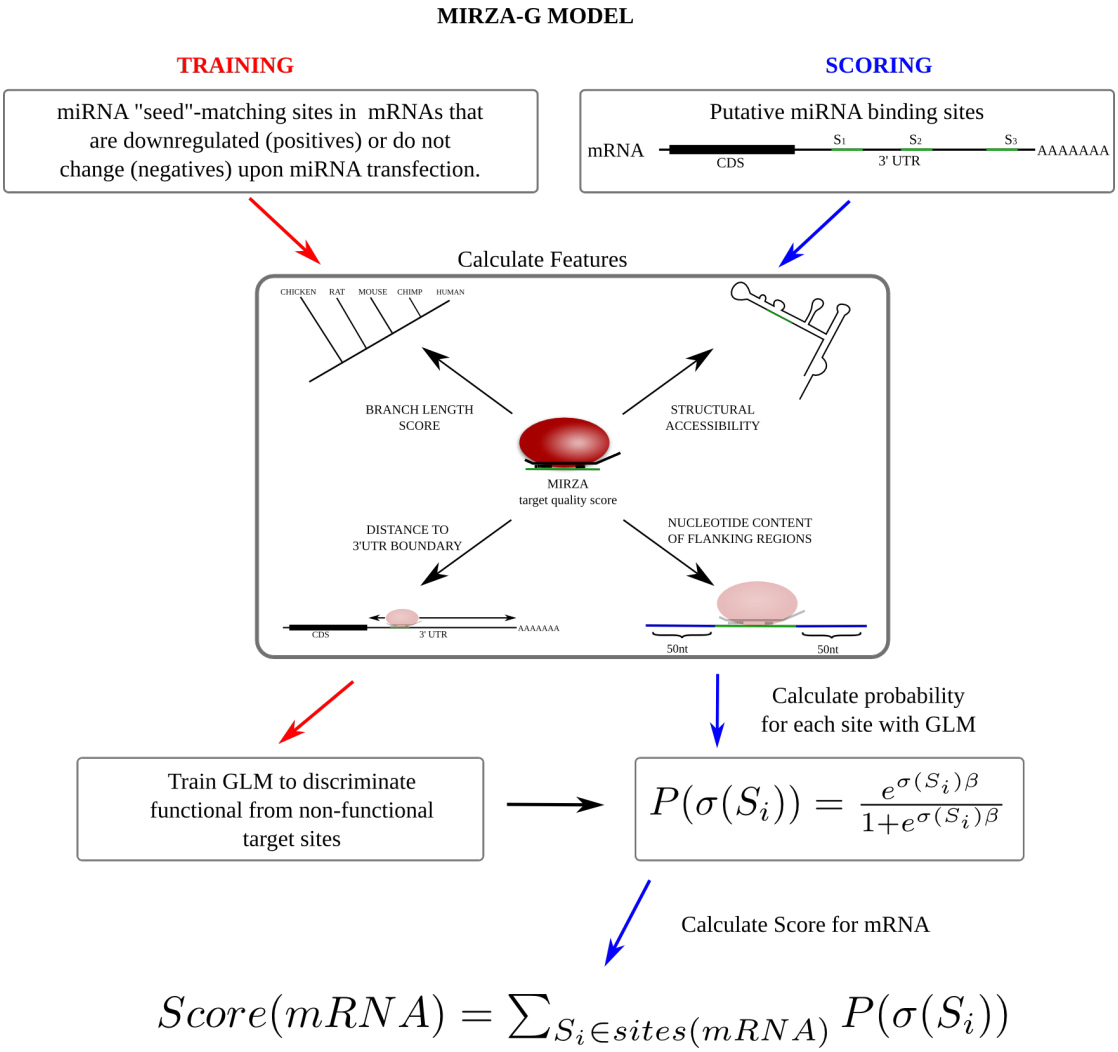


Figure 2.7: Diagram of the approach for predicting miRNA targets with MIRZA-G.

To determine a target quality score threshold for different MIRZA variants we noted that 'canonical' interactions that involve perfect pairing of the miRNA seed have the highest scores with all MIRZA variants. Thus, we employed the procedure that we used before for MIRZA-CLIP [Gumienny and Zavolan, 2015]. That is, with each MIRZA variant, we assigned to each of the 2'998 CLIPed sites from Khorshid et al. [Khorshid et al., 2013] the most likely guiding miRNA. This was the miRNA with the highest target quality score for the site given under the considered MIRZA model. We then predicted the structure of the most likely hybrid between the target site and the guiding miRNA, and divided the sites into

canonical - those with perfect base-pairing over nucleotides 2-8 of the miRNA or perfect pairing over nucleotides 2-7 followed by an adenine (opposite position 1 in the miRNA) - and non-canonical - all other sites. Based on the cumulative distribution of target quality scores for canonical and non-canonical sites, we set a threshold that allowed us to capture the majority of canonical sites without including too many non-canonical sites, that may be artifactually captured. For MIRZA-CLIP a threshold of 50 captures 91% of canonical sites and 18% non-canonical sites, for MIRZA-CHIMERA a threshold of 20 captures 97% canonical and 20% of non-canonical sites, for MIRZA-Class I a threshold of 30 leads to the capture of 94% of the canonical and 18% of non-canonical sites, while for MIRZA-Class IV a threshold of 20 captures 94% of canonical target sites and 20% of the non-canonical target sites.

2.3.3 Wide range of MIRZA quality scores across the targets of a given miRNA

Although we do not focus on this aspect here, it has been proposed that differences in affinity between targets may underlie asymmetries in the crosstalk of mRNAs that bind the same miRNAs [Figliuzzi et al., 2013]. Thus, having shown that the target quality scores computed with MIRZA models correlate very well with the affinities of miRNA-target interactions measured with biochemical methods, we wondered how much variation there is in the affinity of different target sites for a miRNA. Therefore, we determined the MIRZA target quality score for all the sites of all miRNAs that were considered in the genome-wide predictions with MIRZA-G. These had a probability of being functional of at least 0.12 (see [Gumienny and Zavolan, 2015] for details). For each miRNA we have divided the 0 to 10 range of MIRZA target quality scores into bins of 0.2 and have shown the distribution of the target sites of each miRNA as a heat-map, which each line corresponding to a miRNA and the intensity of the color indicating the density of target sites within a bin (Fig. 2.8). It can be seen that the target sites of an individual miRNA span a range of ~ 4 log units or they can differ by ~ 50 fold in the predicted affinity.

2.3.4 Evaluation of the MIRZA models on miRNA transfection data

MiRNAs have been reported to destabilize their mRNA targets, inhibit their translation [Huntzinger and Izaurralde, 2011], and even to increase transcript stability under specific circumstances [Vasudevan et al., 2007]. Of these, perhaps the least controversial is mRNA destabilization, which has been argued to be the dominant mechanism behind the repressive effect of miRNA, with translational repression playing a small, perhaps more transient role [Eichhorn et al., 2014]. The importance of this mechanism is further underscored by observations that miRNA-complementary sites that are conserved in evolution and sites that induce strongest downregulation of their host transcripts upon miRNA transfection have

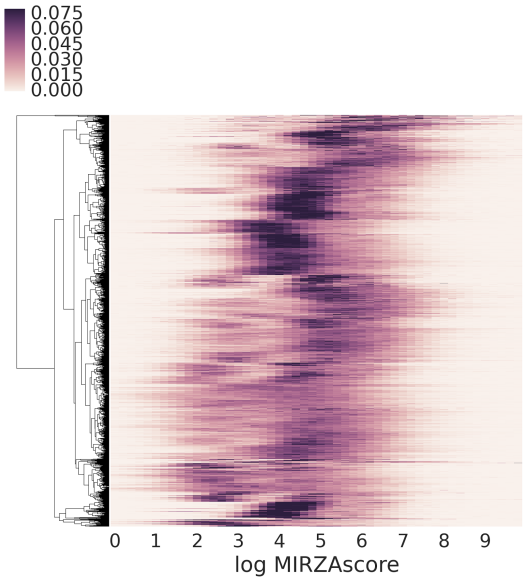


Figure 2.8: Distribution of the MIRZA quality scores of target sites of individual miRNAs. Each line corresponds to one miRNA and the intensity of the color indicates the density of target sites within a particular range of target quality scores, computed with MIRZA-CLIP.

similar properties [Hausser et al., 2009]. Furthermore, acting through the miRNA pathway, small interfering RNAs (siRNA) also destabilize many transcripts (the so-called "off-target" mRNAs) [Jackson et al., 2006]. Thus, it is reasonable to expect that the extent of mRNA destabilization upon miRNA transfection is a robust measure of the strength of interaction between a miRNA and the mRNA. Consequently, the ranking assigned by a computational miRNA target prediction method to mRNAs should correlate well with their change in expression upon miRNA transfection. This is the assumption that we make in discussing the relative performance of various models for miRNA target prediction.

First, we tested whether the models can predict the mRNA expression changes that were induced by individual transfections of miRNAs. To this end, we used data corresponding to 26 miRNA transfections into human cells and one transfection into mouse cells (Table 2). The processing of the transfection data was described extensively in [Gumienny and Zavolan, 2015]. For each type of MIRZA model of miRNA-target interaction we used two variants of the genome-wide MIRZA-G prediction model to predict sites. One of these considered the evolutionary conservation of the sites and the other did not [Gumienny and Zavolan, 2015] (see Fig. 2.7). We sorted targets predicted by each of these models in the

Reference	Data source (Gene Expression Omnibus (GEO) accession / URL)	miRNAs in the data set
Dahiya et al. 2008 [Dahiya et al., 2008]	GSE10150	miR-200c, miR-98
Frankel et al. 2011 [Frankel et al., 2011]	GSE31397	miR-101
Gennarino et al. 2008 [Gennarino et al., 2009]	GSE12100	miR-26b, miR-98
Hudson et al. 2012 [Hudson et al., 2013]	GSE34893	miR-106b
Leivonen et al. 2009 [Leivonen et al., 2009]	GSE14847	miR-206, miR-18a, mir-193b, miR-302c
Linsley et al. 2007 [Linsley et al., 2007]	GSE683	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, miR-16, miR-20, miR-15a, miR-141, miR-200a
Selbach et al. 2008 [Selbach et al., 2008]	http://psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16
Olive et al. 2013 [Olive et al., 2013]	GSE53225	miR-92a

Table 2.2: Data sets of mRNA expression changes following miRNA transfection that were used to test the MIRZA models.

order of their prediction score. We then computed the median log_2 fold-change of the top N predicted transcripts as a function of the number N of top targets considered. The average profiles, computed over the 26 data sets, are shown in Figure 2.9A-B. We found that all four models perform as expected in predicting miRNA targets genome-wide. Consistent with its slightly better performance in predicting the *in vitro*-measured free energy of interaction between miRNAs and target sites, the targets predicted by the MIRZA-CHIMERA model are somewhat more downregulated compared to the targets predicted with MIRZA-CLIP, particularly when the evolutionary conservation of the sites is not taken into account.

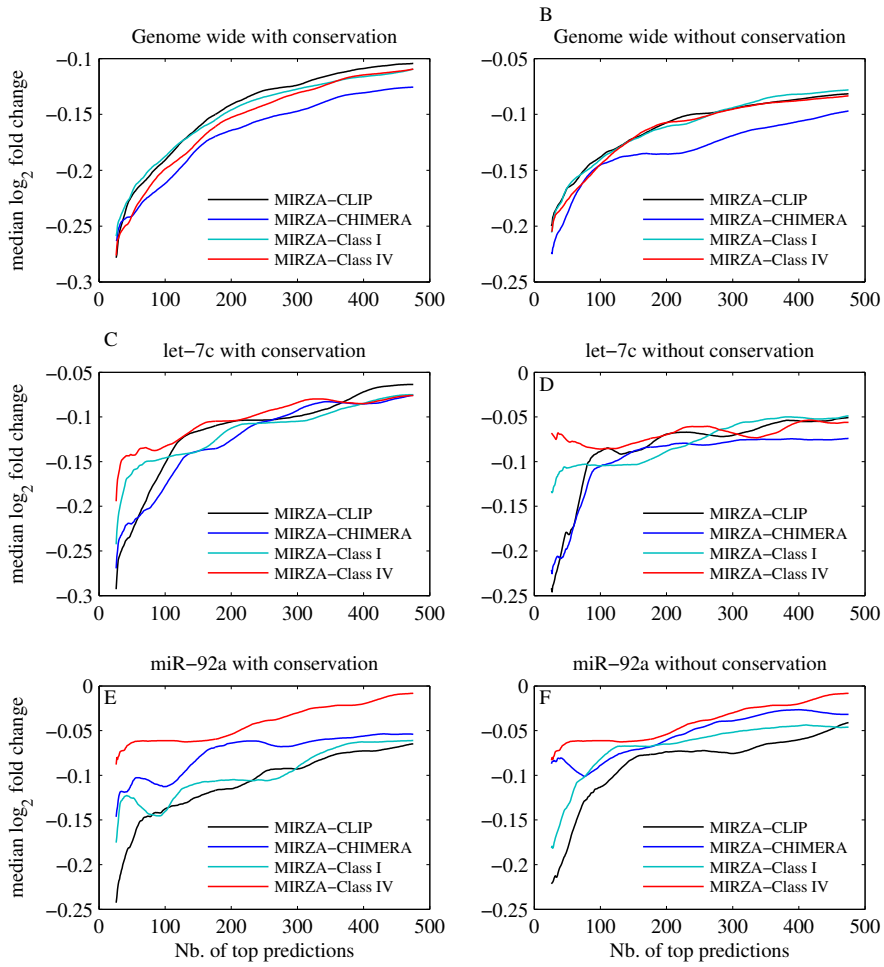


Figure 2.9: Relationship between prediction score and the extent of mRNA downregulation. Genome-wide target predictions were carried out with the MIRZA-G generalized linear model [Gumienny and Zavolan, 2015], within which the target quality scores were calculated with different MIRZA variants: MIRZA-CLIP, MIRZA-CHIMERA, MIRZA-Class I and MIRZA-Class IV. Measurements of mRNA expression in control and miRNA-transfected cells were used to determine the log_2 fold-changes of predicted miRNA targets. (A) Median log_2 fold-change of the top N targets of the transfected miRNA, in function of N, were averaged over a data set of 26 miRNA transfection experiments. (C) Same procedure, but showing the median log2 fold-change of predicted let-7 targets upon let-7 transfection (Table 2.2, data from [Linsley et al., 2007] and [Selbach et al., 2008]) . (E) Same procedure, but showing the median log2 fold-change of predicted targets of the mouse miR-92a upon miR-92a transfection in mouse cells (Table 2.2, data from [Olive et al., 2013]). For (A), (C) and (E), genome-wide predictions were carried out including evolutionary conservation whereas for (B), (D) and (F), without [Gumienny and Zavolan, 2015].

Next we asked whether Class I and Class IV-specific models are more accurate in predicting targets of miRNAs that have been found to yield predominantly class I and class IV chimeras, respectively. As representatives of the first we chose the let-7 family of miRNAs and as a representative of the latter the miR-92a. Because we did not find transfection data for Class IV-chimera forming human miRNAs, we used a data set obtained from mouse cells transfected with the mouse miR-92a. The results, shown in Figure 2.9, panels C-D for let-7 and E-F for miR-92a, clearly indicate that the general MIRZA-CLIP and MIRZA-CHIMERA models are more accurate in predicting transcript downregulation upon miRNA transfection than Class I/IV-specific models. Together with the fact that the sites that are predicted with these models tend to be canonical sites, these results indicate that the origin and relevance of class IV hybrids needs to be further investigated. As mentioned above, a possibility that needs to be ruled out is that the experimental procedure for isolating miRNA-target hybrids via chimeric sequences enriches for non-canonical hybrids that have increased stability prior to ligation.

2.3.5 Inferring a MIRZA model from biochemical data

The results presented above indicate that the MIRZA-CLIP/CHIMERA models explain well both the biochemical data as well as the response of mRNAs to miRNA transfection. However, given the complexity of CLIP experiments and the indirect nature of the resulting data, one wonders whether an even more accurate model of miRNA-target interaction could not be derived from *in vitro* measurements of interaction affinity as obtained in the study of Wee et al. [Wee et al., 2012]. To gain further insight into the design of an efficient experiment, we generated synthetic data sets of hybrids, computed their pseudo-energies of interaction with MIRZA-CLIP, and then asked how our ability to recover the model parameters from the synthetic data sets depends on the number and type of hybrids and the accuracy of the provided pseudo-energies.

First, we simulated the experimental design of Wee et al. [Wee et al., 2012], in which energies of interaction between close variants of a single miRNA (let-7) and their perfectly complementary sequences were measured. There are 1890 possible two point-mutants of let-7, from which we sampled datasets of different sizes. An alternative design is to to measure the energies of interaction between 'random' small RNAs and their partially complementary sequences. In this approach the small RNA is an entirely 'random' sequence whereas the interacting site is a sequence whose complementarity to the small RNA varies. To construct it, we first chose the average number of complementary nucleotides. With probabilities of complementarity chosen uniformly between 0.25 to 1, we can simulate from interactions of random RNA fragments to interactions of perfectly complementary sequences. This second approach is meant to provide datasets containing more information in terms of pairs interacting nucleotides than the first approach. For

both methods, while constructing subsets of various sizes, we aimed to cover uniformly the space of interaction energies and of nucleotide positions involved in the binding. Finally, we considered the possibilities that the measurements are not entirely accurate. To simulate this, we added Gaussian noise to the computed interaction energy for each hybrid with a standard deviation of 0 (no noise), 1%, 5% and 10% of the predicted energy of interaction. For each data set size and each noise level we generated 100 synthetic data sets. To each synthetic data set we applied the simulated annealing procedure that was described in Khorshid et al. [Khorshid et al., 2013] to recover the parameters of the MIRZA model used to generate the pseudo-energies. The results, averaged over the 100 replicates of each setting, are shown in Figure 2.10. They indicate that if the measurement noise is less than 10%, ∼250 hybrids, chosen from across the entire range of expected affinities would be sufficient to recover the model parameters with reasonable accuracy (root mean square difference, RMSD, between recovered and input parameters $<$ 1). If the measurements were very precise (relative error of a few percent), the number of hybrids necessary to recover a model with RMSD $<$ 1 is considerable smaller, ∼100, which is within reach with the technology available today. The experimental design of measuring closely related variants of a single miRNA does not yield equally accurate parameter values from a comparable number of hybrids, presumably due to the limited sampling of nucleotide/position combinations.

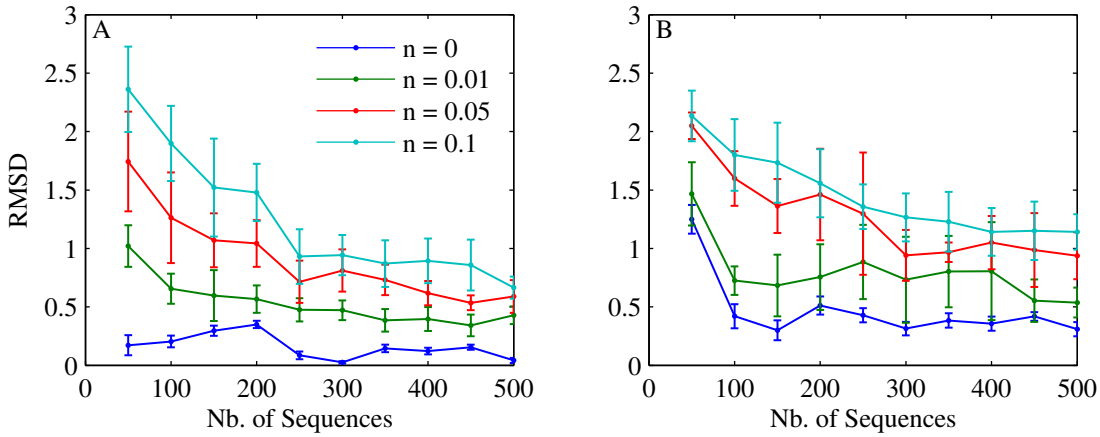


Figure 2.10: Root mean square difference (RMSD) between the MIRZA parameters used to generate the training set and the MIRZA parameters inferred from the training data, as a function of the size of the training set. The colors correspond to the noise added to the training set data (0%, 1%, 5% and 10% of the predicted energy value). For (A), the data sets were generated with the 'randomized' procedure, whereas for (B), the data sets were generated through mutations of the let-7 miRNA.

2.4 Discussion and Perspective

That miRNAs are important for the proper development and function in a large number of species is undisputed. Similar to transcription regulation by transcription factors, miRNA-dependent regulation is 'combinatorial'. That is, a regulator typically has many targets and a target is affected by many regulators. In contrast to transcription factors, miRNAs induce milder changes in target expression, which makes it more difficult to distinguish *bona fide* regulatory effects from biological or experimental variability. Consequently, the field is exploring a number of distinct directions. Many groups have started to explore functional consequences of miRNA-target interaction that go beyond the repression of a single miRNA target into dynamical aspects of the response of a larger network, containing multiple miRNAs and multiple targets [Bosson et al., 2014, Figliuzzi et al., 2013, Bosia et al., 2013, Poliseno et al., 2010, Denzler et al., 2014]. Such a network is quite complex and can exhibit very rich behaviors. For example, a recent study emphasized that even an increased expression of some miRNA targets can be expected in response to the increased expression of a miRNA. This could happen if miRNAs with different efficiencies in target down-regulation compete for the same sites on the target , because over-expression of the miRNA that is less effective in repressing the target could lead to the displacement of the miRNA that is more effective and thus to a net increase in target expression [Nyayanit and Gadgil, 2015]. Additional experiments are necessary to determine whether this behavior occurs *in vivo*.

More generally, given the wide range of behaviors that computational models can predict, it is important to sufficiently constrain them with accurate parameters. Indeed, as described in previous sections, recent studies have started to provide measurements of the concentrations and the rate of interactions between the relevant molecular players. Our work shares this aim. Up to this point we used high-throughput data sets of miRNA binding sites that were derived with various approaches to parameterize a model of miRNA-target interaction. This model allows us to compute the energy of interaction between miRNAs and arbitrary target sites and to carry out genome-wide predictions of miRNA targets. We have shown that the model inferred from sequenced Argonaute/miRNA binding sites predicts quite accurately hybrid energies that are measured with biochemical methods *in vitro*. Furthermore, we have proposed a strategy for deriving a MIRZA-like model from biochemical measurements that can be obtained with the technology available today.

Although on its own, the energy of miRNA-target interaction is not sufficiently predictive of functional interactions, it is one of several informative features that together enable fairly accurate transcriptome-wide predictions. These additional features reflect the secondary structure of the target mRNA, its interactions with RNA-binding proteins, as well as other factors that are yet not understood but can be captured in the degree of evolutionary conservation of the putative miRNA

binding site. Dynamical changes in the miRNA targetome between cell types or cell states will remain difficult to model computationally, but they may be important for the interpretability of experimental data. For example, it has been shown that taking into account tissue/condition-specific isoform expression can improve the prediction of miRNA targets [Nam et al., 2014], because alternative polyadenylation can change the susceptibility of transcripts to miRNA regulation. Conversely, miRNA stability is also subject to regulation, e.g by addition of nucleotides (especially of uridine and adenine) at the 3' end [Kim et al., 2010]. Argonaute protein modifications, mainly phosphorylation, provide another layer of regulation, relieving target repression or changing the subcellular localization [Ha and Kim, 2014]. Nevertheless, the approach that we presented here provides the basis on which more complex, context-specific and even dynamical models describing the impact of miRNA regulation on cellular function can be developed.

Acknowledgements

Jeremie Breda is a Werner-Siemens fellow at Biozentrum and Rafal Gumienny is supported by the Marie Curie Initial Training Network RNPnet project (#289007) from the European Commission. This work was also supported by SystemsX.ch, the systems biology initiative in Switzerland through (RTD StoNets).

Chapter 3

Single cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction

Andrzej J. Rzepiela[1], Souvik Ghosh[1], Jeremie Breda[1], Arnau Vina-Vilaseca[1], Afzal P. Syed[1], Andreas J. Gruber[1], Katja Eschbach[2], Christian Beisel[2], Erik van Nimwegen[1], Mihaela Zavolan[1]

1. Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland
2. Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

Molecular Systems Biology 14 (2018) e8266

Abstract:

MiRNAs are small RNAs that regulate gene expression post-transcriptionally. By repressing the translation and promoting the degradation of target mRNAs, miRNAs may reduce the cell-to-cell variability in protein expression, induce correlations between target expression levels and provide a layer through which targets can influence each other's expression as "competing RNAs" (ceRNAs). However, experimental evidence for these behaviors is limited. Combining mathematical modeling with RNA sequencing of individual human embryonic kidney cells in which the expression of two distinct miRNAs was induced over a wide range, we have inferred parameters describing the response of hundreds of miRNA targets to miRNA induction. Individual targets have widely different response dynamics, and only a small proportion of predicted targets exhibits high sensitivity to miRNA induction. Our data reveal for the first time the response parameters of the entire network of endogenous miRNA targets to miRNA induction, demonstrating that miRNAs correlate target expression and at the same time increase the variability in expression of individual targets across cells. The approach is generalizable to other miRNAs and post-transcriptional regulators to improve the understanding of gene expression dynamics in individual cell types.

Keywords:

miRNA regulation, target downregulation, scRNA-seq, Michaelis-Menten constant, ceRNA

3.1 Introduction

MiRNAs guide Argonaute proteins to mRNA targets, repressing their expression post-transcriptionally [Huntzinger and Izaurralde, 2011]. Measurements of transcript and protein levels following perturbations in the levels of individual miRNAs showed that the fundamental molecular mechanism of mammalian miRNAs is target destabilization, through the recruitment of factors that promote mRNA decay [Lim et al., 2005, Hausser et al., 2009, Guo et al., 2010, Bartel, 2009]. However, time series of mRNA and protein level measurements after miRNA transfection also revealed that repression of target translation precedes the increase in its degradation rate [Bazzini et al., 2012, Hausser et al., 2013, Eichhorn et al., 2014]. A miRNA typically has hundreds of evolutionarily conserved target sites [Lewis et al., 2005, Grün et al., 2005, Gaidatzis et al., 2007], yet only very few predicted targets are down-regulated more than 2-fold in miRNA transfection experiments [Hausser and Zavolan, 2014]. Whereas disruption of miRNA biogenesis impairs the ability of embryonic stem cells to differentiate [Kanellopoulou et al., 2005], and some miRNAs such as the founders of the class, the lin-4 and let-7 miRNA of *Caenorhab-*

ditis elegans have striking developmental phenotypes [Ha et al., 1996, Wightman et al., 1993, Reinhart et al., 2000], most miRNA genes are individually dispensable for development and viability, at least in the worm [Miska et al., 2007]. These observations suggested that strong repression may not be the primary function of miRNAs, and that other functions should be investigated [Ebert and Sharp, 2012].

A computational study of small RNA-dependent gene regulation in bacteria initially proposed that post-transcriptional regulators impose thresholds on the protein levels of their targets in response to transcriptional induction, conferring robustness to transcriptional noise [Levine et al., 2007]. Experiments with target reporters in mammalian systems demonstrated that miRNAs could play a similar role [Mukherji et al., 2011]. Gene expression being a stochastic process, the number of protein molecules expressed from a given gene varies between cells in a cell population. The ratio of variance to mean of the number of protein molecules per cell (the "noise" in protein expression) is predicted to be proportional to the ratio of mRNA translation and mRNA degradation rates [Shahrezaei and Swain, 2008]. Intriguingly, these are the rates that miRNAs modulate so as to decrease protein expression noise. Indeed, a recent study reported increased variability in CD69 protein expression across miRNA-deficient, developing mouse thymocytes [Blevins et al., 2015]. However, as the reduction in target protein noise is predicted to scale as the square root of the miRNA-induced change in protein level [Osella et al., 2011, Schmiedel et al., 2015], which is small for the vast majority of evolutionarily conserved miRNA targets [Hausser et al., 2013, Eichhorn et al., 2014, Hausser and Zavolan, 2014], it is unlikely that many of the predicted miRNA targets are regulated in this manner.

It has also been proposed that at the cellular level, miRNAs provide a "channel" through which the many predicted miRNA targets "communicate" as "competing RNAs" (ceRNAs) [Poliseno et al., 2010, Figliuzzi et al., 2013, Cesana et al., 2011, Karreth et al., 2015, Wang et al., 2013]. Rough estimates of the number of potential binding sites for a miRNA (also called miRNA "target abundance") are in the range of $\sim 10^5$ sites per cell, much higher than the number of cognate miRNA molecules [Denzler et al., 2014]. In this regime, where the targets are already in high excess relative to the miRNAs, over-expressing a single target could not appreciably affect the expression of the other targets. Yet examples of ceRNAs continue to emerge [Poliseno et al., 2010, Figliuzzi et al., 2013, Cesana et al., 2011, Karreth et al., 2015, Wang et al., 2013]. These estimates of target abundance did not consider the possibility that targets may not be equivalent in their ability to bind and sequester miRNAs. Indeed, a computational analysis suggested that miRNA targets have asymmetric relationships, high affinity targets being able to sequester miRNAs from low affinity targets, at comparable target concentrations, but not the other way around [Figliuzzi et al., 2013]. In vitro measurements indicate that miRNA target sites can have widely different affinities for the miRNA-Argonaute complex [Wee et al., 2012], an observation that is supported by measurements of

Argonaute dwelling times on individual miRNA target sites [Chandradoss et al., 2015]. However, estimates of *in vivo* miRNA-target interaction constants are lacking.

Taking advantage of a system in which the expression of a single miRNA precursor can be induced over a wide concentration range, we measured the transcriptomes of thousands of individual cells and assessed how the expression levels of miRNA targets relate to the expression level of the miRNA. We obtained experimental evidence for behaviors that were previously suggested by computational models or evaluated only with miRNA target reporters. These include the non-linear, ultrasensitive response of miRNA targets to changes in the miRNA concentration as well as the dependency of the variability in target levels between cells on the concentration of the miRNA. Furthermore, we found that only a small fraction of predicted targets are highly sensitive to changes in miRNA expression. With a computational model we illustrate how these targets can influence the expression of other targets as competing RNAs. Our approach is applicable to other post-transcriptional regulators of mRNA stability, allowing the analysis of their concentration-dependent impact on the transcriptome.

3.2 Results

3.2.1 A system to study the impact of miRNA expression on the transcriptome of individual cells

MiRNA target reporters are widely used to study miRNA-dependent gene regulation. However, these reporters are often expressed at much higher levels than when expressed from their corresponding genomic loci. Furthermore, these reporters do not respond to the regulatory influences to which the endogenous transcripts respond. To circumvent these issues and investigate the crosstalk of miRNA targets in their native expression context, we used a human embryonic kidney (HEK) 293 cell line, i199 [Hausser et al., 2013], in which the expression of the hsa-miR-199a miRNA precursor and of the green fluorescent protein (GFP) can be simultaneously induced by doxycycline, from a pRTS1 episomal vector (Figure 3.1A). To assess the reproducibility of the inferred sensitivity parameters for miRNA targets, we used a related cell line, i199-KTN1 [Hausser et al., 2013], derived from i199 through the stable integration of a target of hsa-miR-199a-3p. This target consisted of the renilla luciferase coding region followed by the 3' untranslated region (UTR) of kinectin 1 (KTN1). We reasoned that these similar but not identical cell lines should allow us to assess the reproducibility of the inferred parameters, which we do expect to vary between more distant cell types due to differences in the expression of regulatory factors.

The processing of hsa-mir-199a gives rise to two mature miRNAs, hsa-miR-199a-5p and hsa-miR-199a-3p. These miRNAs have distinct "seed" sequences (at

positions 2-7 of the miRNA) and therefore, largely non-overlapping target sets; only 7 of the top 100 targets predicted [Gumienny and Zavolan, 2015] for each miRNA are shared. The bidirectional nature of the promoter in the pRTS1 vector was characterized before, by fluorescence-activated cell sorting [Bornkamm et al., 2005]. In our construct, the luciferase protein-coding sequence has been replaced by a pri-miRNA. As no method is currently available for simultaneously measuring the expression of a miRNA and of a protein-coding gene in single cells, we assessed whether the two bi-directionally transcribed RNAs have correlated expression in cell populations. Indeed, by RT-PCR we found that the expression of hsa-miR-199a-5p and of the GFP mRNA, in cell populations induced with different concentrations of doxycycline, were highly correlated (Figure 3.1B, Spearman $r = 0.91, p = 1.74 \cdot 10^{-7}$). Furthermore, the expression of both mature miRNAs processed from the hsa-mir-199a precursor increased in parallel to the concentration of the inducer, as expected (Appendix Figure 3.6A). Altogether, these data indicates that the level of GFP mRNA can serve as a "proxy" for the miRNA levels in studying the response of miRNA targets to the miRNA in individual cells. Carrying out Argonaute 2 protein crosslinking and immunoprecipitation in fully induced ($1\mu g/ml$ doxycycline) HEK 293 cells, we confirmed that hsa-miR-199a-5p and hsa-miR-199a-3p were incorporated into the miRNA effector complex, and were among the highest represented miRNAs (Appendix Figure 3.6B).

We then induced cells with doxycycline concentrations spanning the 0 $\mu g/ml$ to 1 $\mu g/ml$ range (as described in 3.4), then pooled the cells and carried out mRNA 3' end sequencing of 3280 distinct i199 and 3143 i199-KTN1 cells, on a 10x Genomics platform. In parallel, we carried out bulk mRNA sequencing from both non-induced i199 cells and cells that were fully-induced (1 $\mu g/ml$ doxycycline). The distribution of the number of distinct transcripts obtained from individual single cells is shown in Appendix Figure 3.6C. GFP mRNAs were captured from 43% of the i199 cells, in which the mean GFP mRNA expression was 32 transcripts per million (TPM) (Figure 3.1C). mRNA expression levels inferred either by averaging over single cells (SC) with no GFP mRNA or from bulk sequencing of non-induced cell populations (CP), were highly correlated (Spearman r of $\log_2$ expression values $= 0.89$, $p < 10^{-15}$, Figure 3.1D). The expression of the top 100 MIRZA-G-C-predicted targets of the two miRNAs [Gumienny and Zavolan, 2015] was significantly lower in cells with high GFP mRNA expression (> 6.8 TPM) compared to cells with no GFP expression (0 TPM, Figure 3.1E). Importantly, the expression of predicted targets decreased in parallel with the increase in GFP mRNA levels, further indicating that the GFP mRNA is a good proxy for the miRNA expression in individual cells (Appendix Figure 3.6D,E). Finally, the miRNA-induced changes in target expression, inferred either from bulk or from single cell sequencing of strongly induced and uninduced cells, were significantly correlated (Figure 3.1F). The results were reproduced in the related i199-KTN1 cell line (Appendix Figure 3.7). Results of a parallel analysis with miRNA targets

predicted by TargetScan [Garcia et al., 2011] are shown in the Appendix Figures 3.11-3.17. Altogether, these results indicate that the system behaves as expected and can be used for further analysis of miRNA-dependent gene regulation in single cells.

3.2.2 miRNA targets follow theoretically predicted behaviors in response to miRNA induction

The dynamics of small networks composed of miRNAs and targets has been investigated computationally, with stochastic models [Bosia et al., 2013,Figliuzzi et al., 2013]. Bosia et al [Bosia et al., 2013] predicted that the coefficient of variation (C_V) of miRNA targets increases with the transcription rate of the miRNA, showing a local maximum in the region where the miRNA and targets are in equimolar ratio. The correlation of expression levels of mRNAs that are targeted by the same miRNA was predicted to exhibit a maximum around the same threshold. We used a similar simple model of miRNA-dependent gene regulation to predict the behavior of targets in our experimental system. Briefly, we considered M mRNA targets of a given miRNA, each with a specific transcription rate α_i, decay rate δ_i, and level m_i, with $i \in \{1, ..., M\}$. Target i could bind a miRNA-containing Argonaute (Ago) complex at rate k_{on} and dissociate from the complex at rate k_{off}. Because in our experimental system we induced miRNA expression to specific stable levels before carrying out the mRNA sequencing, we neglected the dynamics of the miRNA and assumed that the total number A of Ago-miRNA complexes in a given cell was constant, though varying between cells. The number of free Ago-miRNA complexes is then given by $A_F = A = \sum_{j=1}^{M} A_j$. Finally, we assumed that Ago-miRNA-bound mRNAs decay at rates k_{cat}. Under this simple model (see also [Hausser and Zavolan, 2014]), free mRNAs (m_i) and miRNA-bound mRNAs (A_{m_i}) follow the dynamics described by the system of $2M$ differential equations

$$\frac{\partial m_i(t)}{\partial t} = \alpha_i - \delta_i m_i(t) - k_{on_i} m_i(t) \left(A - \sum_{j=1}^{M} A_{m_j}(t) \right) + k_{off_i} A_{m_i}(t) \quad (3.1)$$

$$\frac{\partial A_{m_i}(t)}{\partial t} = k_{on_i} m_i(t) \left(A - \sum_{j=1}^{M} A_{m_j}(t) \right) - k_{off_i} A_{m_i}(t) - k_{cat_i} A_{m_i}(t) \quad (3.2)$$

We carried out stochastic simulations of a system with four miRNA targets (Figure 3.2A), choosing parameters of target expression and interaction with the miRNA such that (1) target expression spanned a broad range, (2) they underwent miRNA-dependent down-regulation at either low (targets **a** and **b**), or high (targets **c** and **d**) miRNA levels, and (3) down-regulation of all targets was moderate, as generally observed in experiments. The response of individual *in silico* targets to miRNA induction is shown in Figure 3.2A. Figures 3.2B,C show the variability of

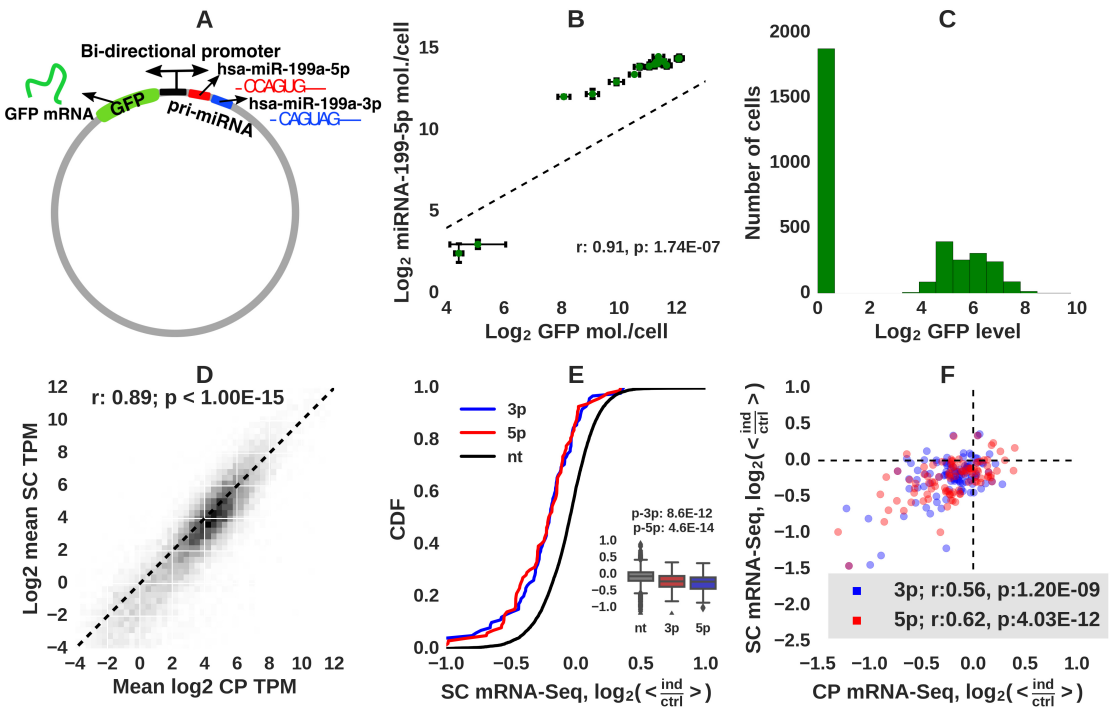


Figure 3.1: **Design and characterization of the experimental system. A.** Schematic representation of the construct used to express hsa-miR-199a-5p (red), hsa-miR-199a-3p (blue), and the reporter GFP mRNA from a bidirectional promoter. Shown are also the "seed" sequences (nucleotides 2-7) of the two miRNAs. **B.** The expression levels of hsa-miR-199a-5p and GFP mRNA, measured from cell populations by quantitative PCR, are highly correlated. **C.** Histogram of normalized GFP mRNA expression (TPM) in individual i199 cells. **D.** Correlation of mRNA expression levels estimated from SC sequencing (1875 T^0 cells (see text for definition) from which no GFP mRNA was captured) and from CP mRNA-seq (6 replicates of non-induced cell populations). **E.** Cumulative distribution of expression differences of the top 100 targets of hsa-miR-199a-5p (red), top 100 targets of hsa-miR-199a-3p (blue), and of 7347 remaining, "background" genes (black) between cells expressing highest and lowest GFP levels [216 T^∞ cells with > 6.8 TPM GFP ("ind") vs. 1875 T^0 cells with 0 TMP GFP ("ctrl")]. Box plots of $\log_2$-fold change of non-targets, top 100 miR-199a-3p and top 100 miRNA-199a-5p targets are shown in the inset. P-values of the rank-sum test comparing targets and non-targets are also shown. Horizontal line is a mean, box shows where 50% of data points are (interquartile range, IQR), whiskers show points within 1.5·IQR from 25/75-percentile border of the box. **F.** Scatter plot of expression differences of the top 100 targets of each miRNA, estimated from bulk sequencing (CP) or from single cell sequencing (T^∞ and T^0 cells defined as for previous panel).

target expression between simulated cells and the pairwise correlations of target expression levels across all simulated cells, as functions of the total miRNA level. Similar to the predictions of Bosia et al. [Bosia et al., 2013], the targets in our *in silico* system also experience destabilization, increased correlation and increased expression noise, all within a limited range of miRNA expression, i.e. at a specific threshold. Figure 3.2B also shows that for each target, the coefficient of variation increases in function of miRNA expression level, as the target expression level is reduced by the miRNA, and that targets with low expression level have higher coefficients of variation compared to highly expressed targets. Furthermore, there is a noticeable spike in the coefficient of variation of each target, in the region where the target experiences a hypersensitive down-regulation in response to the miRNA (see also Appendix Figure 3.7A). The miRNA also induces correlated changes in its targets (Figure 3.2C); targets with high sensitivity to the miRNA, that are repressed at low miRNA concentrations (**a** and **b** in our example), exhibit the highest correlation coefficient, and over a widest range of miRNA concentrations. However, targets that differ strongly in concentration of the miRNA that triggers their response, or in the magnitude of miRNA-induced decay appear uncorrelated (**c** with respect to the others in our example).

We then turned to the experimental data. For both miRNAs and both cell lines, the total target level (see section 3.4 for target selection) exhibited the expected threshold-decrease in function of the GFP expression level, which we used as proxy for the miRNA expression (Figure 3.2D,G and Appendix Figure 3.7D). The C_V and r_P values, computed as ratios to the corresponding values for a similarly-sized set of non-targets also showed the expected behaviors. Namely, the coefficient of variation in total target expression across individual cells increased with the GFP expression (Figure 3.2E,H, see also Appendix Figure 3.7B,E), while the mean pairwise correlation coefficient of target expression in individual cells peaked at an intermediate level of GFP mRNA expression (Figure 3.2F,I, see also Appendix Figure 3.7C,F). Randomizations showed that in spite of the large noise, indicated by the size of the error bars, the C_V of targets remained larger than that of non-targets and the correlation of target expression larger than the correlation of non-target expression. Thus, even though the noise of single cell experiments is large and mRNA capture incomplete, the experimental data follow the theoretical predictions and the simulations.

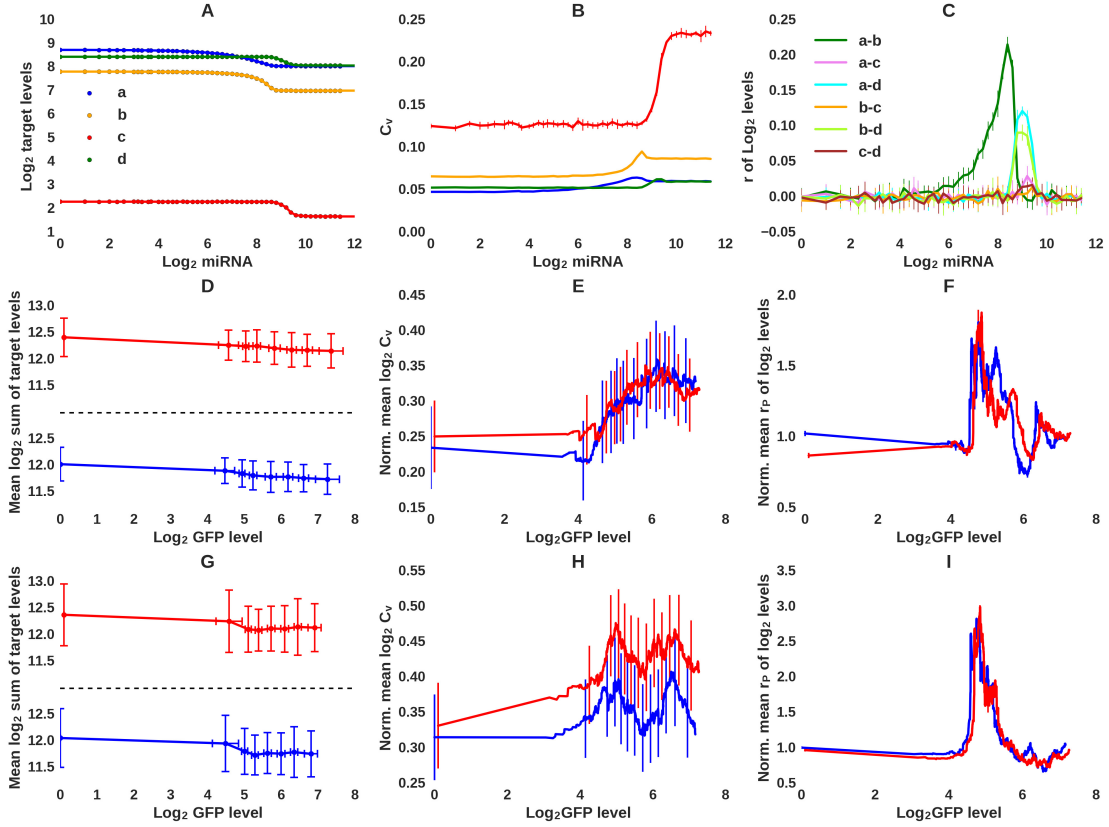


Figure 3.2: **Expected and observed response of miRNA targets to miRNA induction in single cells. A.** Results of numerical integration (Eqs (3.1)-(3.2), solid lines) and the average of six stochastic simulations (dots) of a model with four target genes (indicated by distinct colors) chosen to cover a wide expression range and to have either high or low sensitivity to the miRNA. Fifty *in silico* cells, each with a defined miRNA concentration were simulated. **B.** Coefficient of variation (C_V) of *in silico* target levels across cells, calculated in function of miRNA expression, from the simulation trajectories. **C.** Pearson's correlation coefficients of expression levels of pairs of genes from in silico cells, calculated in function of miRNA expression from the simulation trajectories. **D,G.** Total expression ($\log_2$ sum of TPMs) of 100 lowest A_F^C hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets (see also Materials and Methods for target selection) in the i199 (**D**) and i199-KTN1 (**G**) cells, in function of $\log_2$ GFP expression in the same cells. **E,H.** Mean C_V and **F,I.** Mean Pearson's pairwise correlation coefficients for miRNA targets in function of GFP expression in i199 (**E,F**) and i199-KTN1 (**H,I**) cells. Averages were calculated from the two hundred cells with GFP expression closest to a specific expression level. C_V values are shown as ratios to corresponding values computed for all non target mRNAs (**E,H**) and r_P to mean of 50 evaluations of random selection of 100 non-target genes (**F,I**). For **B,C,D** and **G** plot standard deviations are shown, for **E,F,H** and **I** standard error.

3.2.3 The sensitivity of individual targets to miRNA regulation can be inferred from their expression in cells with varying miRNA level

We used the computational model described in Eqs (3.1),(3.2) to derive two measures of target sensitivity to miRNA regulation. First, we derived the Michaelis-Menten-like constant [Wee et al., 2012] $K_{M_i} = \frac{k_{off_i}+k_{cat_i}}{k_{on_i}}$ defined as the ratio of the dissociation rate of mRNA i from the miRNA-primed Argonaute protein (whether or not accompanied by Ago-catalyzed decay) and their rate of binding. We further derived the level of free Ago-miRNA complexes at which a specific target i is halfway between its maximum level T_i^0, realized when the miRNA is not expressed, and its minimum level T_i^∞, realized when the miRNA is in high excess relative to all targets. As shown in the Methods, this critical concentration is given by

$$A_{F_i}^C = \frac{K_{M_i}}{T_i^0/T_i^\infty} \tag{3.3}$$

We then devised a procedure for inferring these two parameters for each miRNA target from the experimental data, which has a high level noise (total target levels vary almost 2-fold in individual cells with similar GFP expression (Figure 3.2D,G), for reasons that may include the low mRNA capture rate and the imperfect coupling of miRNA and GFP mRNA levels). We used the system described in Eqs. (3.1),(3.2) to test procedures for analyzing noisy single cell data such that we can infer target-specific parameters at the limit of accuracy afforded by the single cell experiments. Relevant for the inference are the expression levels of targets in the absence of the miRNA, the expression levels when the miRNA is present at maximal concentration, and the expression levels in all cells in which the miRNA has intermediate expression. Thus, we generated *in silico* data with the computational model (Figure 3.3A), added noise in target levels comparable to the noise observed *in vivo* (Figure 3.3B) and then experimented with the selection of cells to use in the inference and with the smoothing of the target levels (Figure 3.3B,C) to most accurately recover the input parameters (see also Methods and Appendix Figure 3.9). In particular, different miRNA targets respond at different miRNA concentration, and only cells in which the miRNA concentration is in the relevant range for that target could be used for inferring the shape of the target's response. Thus, to select cells that are relevant for the inference of parameters of all targets in parallel, we examined the dependence of average target level as a function of the miRNA concentration in a cell. As even the average target level varies quite widely between cells with similar miRNA concentration (Figure 3.3B), we explored procedures for smoothing average expression levels as a function of miRNA expression before the selection of cells for the inference, as described in Methods. The region of target sensitivity to the miRNA is indicated by the red line in Figure 3.3B, and the gradient in mean target level as a function of miRNA

concentration is shown in Figure 3.3C. The free miRNA levels inferred from these *in silico* data showed that only when the total miRNA level is sufficiently high to occupy all the available target sites (Figure 3.3D) does free miRNA accumulate, as expected. The correlation between target-specific input and recovered parameters (Figure 3.3E, Pearson's $r = 0.56$, p-value $= 3.0 \cdot 10^{-20}$) was at the upper bound set by the level of noise in the simulated data, as shown by correlation between the parameters recovered from two simulations that only differed in the measurement error added to the target expression levels in the simulated cells (Figure 3.3F).

3.2.4 A limited number of targets exhibit high sensitivity to miRNA induction

We then turned to estimating the sensitivities of the predicted miRNA targets from the experimental data. For each miRNA, we selected the 300 MIRZA-G-C-predicted targets with highest prediction scores [Gumienny and Zavolan, 2015], that had an expression level of at least ∼8 TPM when the miRNA was not expressed, and underwent at least 8% downregulation at the highest miRNA concentration ($\log_2 T_i^\infty / T_i^0 < -0.12$). We used cells with $\log_2$ GFP expression of 0 TPM (1875 and 1629 cells for i199 and i199-KTN1 cells, respectively) to infer target levels T_i^0, when the miRNA is not expressed, those with more than 6.8 TPM GFP (216 cells for i199 and 205 for i199-KTN1) to infer target levels T_i^∞ at saturating miRNA concentration, and all other cells to construct the $\tilde{T}$ matrix of individual target expression levels in single cells with intermediate miRNA expression. Applying the inference described in the methods (3.4), we obtained A_F^C (Figure 3.4A) and K_M (Figure 3.4B) parameters for all targets, and found that their distributions covered a 4 to 8 fold range. The average response of the 20 targets with lowest and highest values of these two parameters to miRNA induction is shown in Figure 3.4C. For both hsa-miR-199a-5p and hsa-miR-199a-3p miRNAs, target parameters inferred independently from the two cell lines were significantly correlated (Figure 3.4D,E), indicating the robustness of our results. For hsa-miR-199a-3p, Pearson's correlation coefficients were 0.49 (p-value $= 2.6 \cdot 10^{-11}$) for A_F^C, and 0.4 (p-value $= 5.9 \cdot 10^{-8}$) for K_M, while for hsa-miR-199a-5p they were 0.43 (p-value $= 3.6 \cdot 10{-}8$) for A_F^C, and 0.34 (p-value $= 2.2 \cdot 10^{-5}$) for K_M. Especially apparent on the scatter plot of A_F^C values is a small group of targets that respond at low miRNA concentrations and thus have low A_F^C in both cell lines (see also Figure 3.4C). These low A_F^C targets have higher prediction scores and are enriched in DNA-binding factors compared to the high A_F^C (Appendix Table 3.1 and 3.2 and Appendix Figure 3.10A). The measure that is most broadly used to validate computational target predictions is the change in expression that predicted targets experience upon strong miRNA induction (Figure 3.4F). Sorting targets by their MIRZA-G-C scores and computing the average fold change (between cells with high (T_i^∞) and no (T_i^0) miRNA expression) of the top x targets as a function

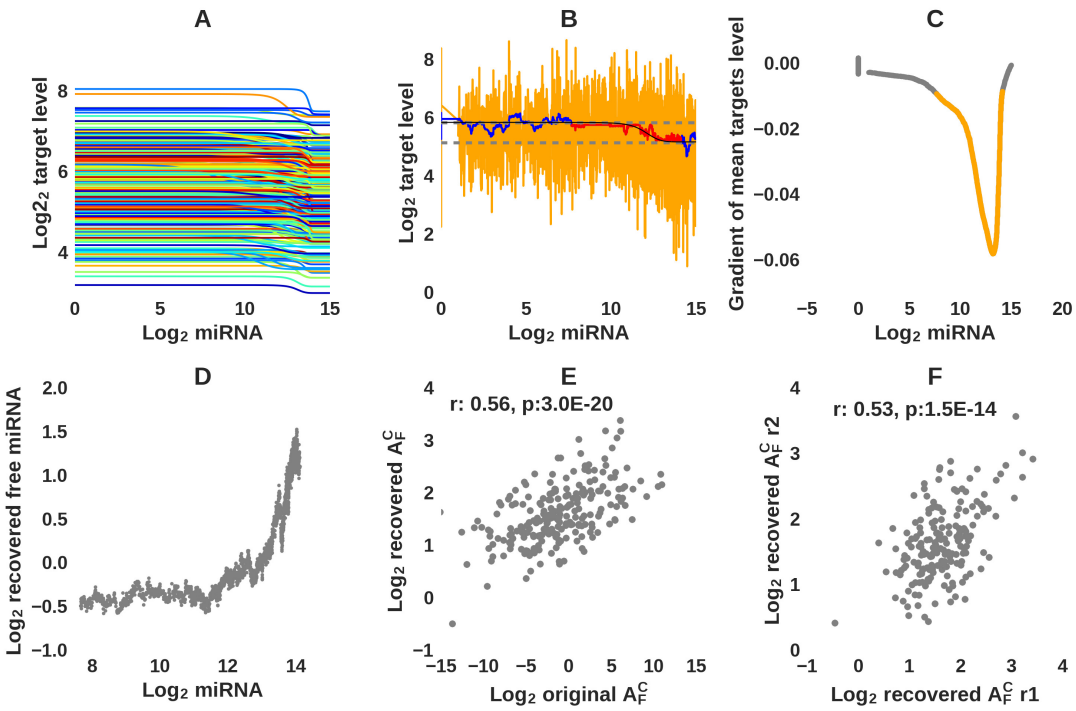


Figure 3.3: **Validation of the approach for inferring target sensitivity from single cell data. A.** Response of 300 *in silico* targets, each with associated parameters describing its transcription, decay, rates of binding to and dissociating from the miRNA (values drawn from distributions around experimentally measured values, see Appendix Figure 3.9) in response to increasing miRNA concentration. **B.** Noise (orange) was added to the target expression (black) and then running means (blue) were calculated over increasingly wider windows to ensure that the estimated expression levels T_{ij} for gene i in cell j (for cells used in the inference (red)), were between the maximum (T_i^0) and minimum (T_i^∞) levels, corresponding to no miRNA being expressed and to the miRNA being expressed at very high levels in the cell (allowing for a small tolerance c; dashed lines). **C.** Cells for which the gradient of the total target level with respect to the miRNA level was less than -0.01 (shown in orange, and corresponding to the points shown in red in panel (**B**)) were used to construct the $\tilde{T}$ matrix of gene expression levels per cell. **D.** Scatter plot of the total miRNA levels that were used as input to the model and the levels of free miRNA inferred from the simulated data. **E.** Scatter plot of the input vs. inferred A_F^C values. The Pearson correlation coefficient and its associated p-value are also shown. **F.** Scatter plot of A_F^C values inferred from *in silico* data that were generated with the same input target parameters, but to which two distinct sets of "measurement errors" were applied. The Pearson correlation and associated p-value are also shown.

of x, we indeed found that the highest-scoring targets undergo the largest down-regulation (Figure 3.4F, dotted lines), as expected. Similar patterns of stronger down-regulation of top targets was also apparent when we sorted targets based on their sensitivity to the miRNA reflected in the A_F^C parameter (Figure 3.4F, dashed lines). However, the best indicator of the degree of down-regulation of a predicted target was its inferred K_M (Figure 3.4F, full lines). This could indicate that the inferred K_M values are dominated by k_{cat}, the rate of target degradation when complexed to the miRNA, while the rates of miRNA-target association and dissociation vary less between targets. Finally, we examined what features of the predicted miRNA binding site were most informative for the A_F^C, K_M, and fold-change of the target (Appendix Figure 3.10B). For this, we selected only the 231 targets with a single binding site (for either of the miRNAs), to ensure that the site context effects could be attributed unambiguously. Consistent with prediction models being trained to predict mRNA level changes upon miRNA transfection, the prediction scores correlate best (in absolute value) with the fold-change of the predicted targets in cells with high miRNA expression compared with low miRNA expression. Measures related to the A/U content in the vicinity of sites and their relative location in 3' UTRs are most predictive for A_F^C and K_M, whereas the degree of evolutionary conservation is most informative for the fold-change of the target.

3.2.5 Implications for the ceRNA function of miRNA targets

To evaluate the implications of our results for the debate about the prevalence of competing endogenous RNAs [Denzler et al., 2014, Bosson et al., 2014], we used again our computational model with realistic K_M values and explored the effect of one miRNA target (the ceRNA) on the expression of all other targets. Target parameters were set as described in section "*In silico* analysis", to cover the range inferred from various experimental systems. We note that a ceRNA is only one species of RNAs expressed in a cell and, for the vast majority of parameter values that are in the range determined for other RNAs in the cell, the ceRNA is predicted to cause expression changes that are very low, below 1%. Nevertheless, we illustrate some of the more interesting scenarios below. We set the decay rate of the free ceRNA to $0.1/h$, its $k_{on} = 0.2/h$, similar to those of other targets, and we varied the k_{off} and k_{cat} to achieve either low or high K_M. We then asked how much the expression of the pool of targets with either low (less than $0.02M$) or high (greater than $2M$) K_M targets changes, when the ceRNA is expressed at different levels. As shown in Figure 3.5, we found that highly expressed ceRNAs with low K_M can induce the upregulation of low, and especially high K_M targets. However, substantial upregulation of other targets, larger than a few percent, is only achievable when the ceRNA has very high transcription rate and does not decay when in complex with the miRNA. This is what one

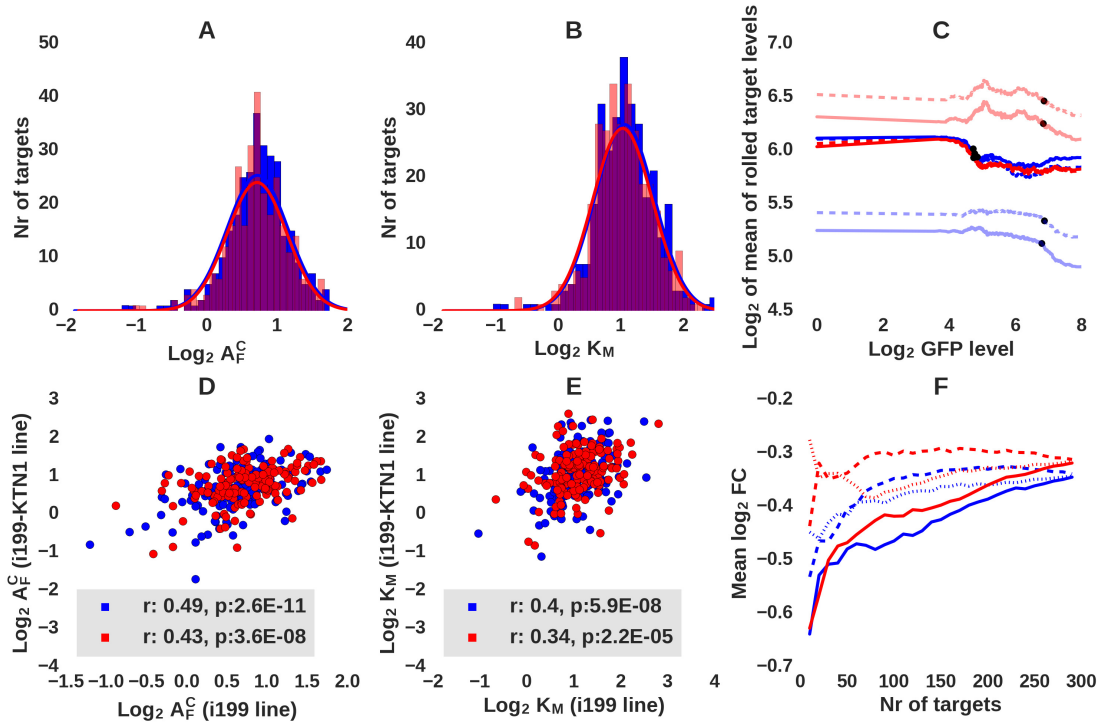


Figure 3.4: **Parameters describing the response of individual targets to changes in miRNA expression.** Histograms of A_F^C (**A**) and K_M (**B**) values of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets, inferred from the i199 cell line. The lines indicate the best-fitting Gaussian distributions. **C.** Response of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets to the miRNAs in i199 cells. Targets were selected based on A_F^C (dashed lines) or K_M (full lines) values, targets with low values of the respective parameters are shown in strong color and those with high values in faded colors. 20 targets were summed up for each category. Dots show the point where the targets have undergone 1/2 of their maximal down-regulation. **D.** Scatter plot of $\log_2 A_F^C$ values inferred for individual targets from the i199 and i199-KTN1 cell lines. Shown are also Pearson's correlation coefficients and corresponding p-values. **E.** Scatter plot of $\log_2 K_M$ values inferred for individual targets from the i199 and i199-KTN1 cell lines. Shown are also Pearson's correlation coefficients and corresponding p-values. **F.** Average $\log_2$ fold change of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets as a function of the number of top targets considered, where predictions are made based either on K_M values (highest to lowest, full lines), A_F^C values (lowest to highest, dashed lines) or MIRZA-G-C scores (highest to lowest, dotted lines).

intuitively expects, namely that a ceRNA can influence the expression of other targets when its expression is comparable to that of all other targets taken together. On the other hand, if the ceRNA has high K_M, its influence on the expression of

other targets will be negligible. These results strongly suggest that ceRNAs that were observed so far are highly expressed transcripts that are relatively resistant to degradation. These would be able to "sponge" miRNAs from targets which the miRNA strongly destabilizes, these having high k_{cat} and high K_M. Good candidates seem to be the relatively recently described circular RNAs [Memczak et al., 2013, Hansen et al., 2013]. However, given the multiple constraints that a transcript has to fulfill to be able to function as a ceRNA (very high transcription and/or stability, low K_M), this mode of regulation should be rare.

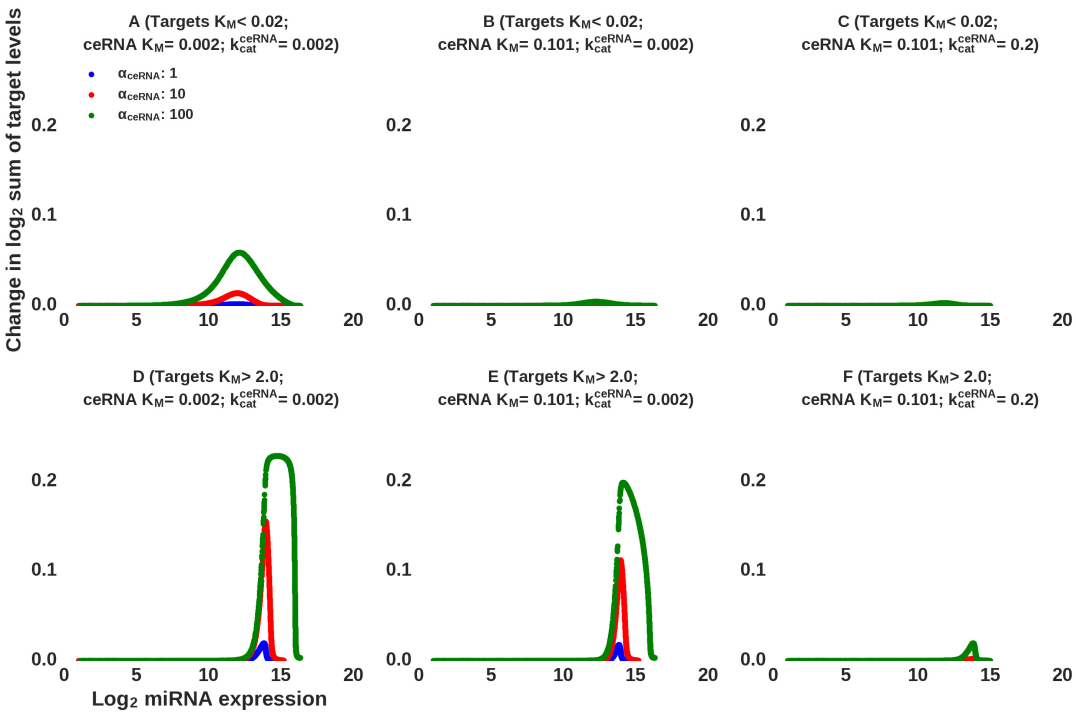


Figure 3.5: **Predicted response of different types of miRNA targets to the induction of a ceRNA.** A competing RNA with low (**A,D**), or medium (**B,C,E,F**) K_M is transcriptionally induced at three different levels. **A-C** show the response of targets with low $K_M(< 0.02)$ to the transcriptional induction of the ceRNA whereas **D-F** show the response of targets that have high K_M (> 2.0). The decay rate of the ceRNA when unbound to the miRNA δ_{ceRNA} was set to 0.1/h, whereas when bound to the miRNA the ceRNA was assumed to be either stabilized and long-lived (k_{cat} was set to 0.002/h) (**A,B,D,E**) or destabilized and shorter-lived (k_{cat} was set to 0.2/h) (**C,F**).

3.3 Discussion

Single cell RNA-sequencing has opened a new route to the quantitative understanding of cellular functions. This technology has been used to characterize transcript isoforms and gene expression [Shalek et al., 2013], to improve classification of cell types [Buettner et al., 2015], and to discover new, particularly rare types of cells [Grün et al., 2015]. The relatively low rate of mRNA capture and the large technical noise remain important issues for single cell sequencing, particularly with droplet-based methods, which rarely use spike-ins for normalization [Ziegenhain et al., 2018, Gao, 2018]. However, developments such as unique molecular identifiers [Grün et al., 2014] push the boundary towards ever increasing accuracy. Although data analysis methods are still in flux, in our study, we used known properties of miRNA targets to gage whether our processing of the data is appropriate. For example, we showed that miRNA target down-regulation computed from the inferred target levels in single cells is similar to the down-regulation inferred from bulk sequencing. Single cell analysis has also been used to infer parameters of gene expression (see [Munsky et al., 2015] for a recent review).

Although it was proposed that miRNAs buffer stochastic fluctuations in gene expression between cells [Hornstein and Shomron, 2006], experimental data pertaining to expression of miRNA targets in individual cells with different levels of miRNA expression is very limited. Some studies estimated the effect of endogenous miRNAs on the protein expression noise of target reporters with multiple miRNA-complementary sites [Mukherji et al., 2011, Schmiedel et al., 2015]. The reduction in protein expression noise has been related to the degree of miRNA-induced down-regulation, which is generally limited, except for reporters that carry multiple perfectly-complementary miRNA binding sites in their 3' UTRs. Additional studies are needed to evaluate the extent to which miRNAs regulate the expression noise of their targets in their native context (see also [Schmiedel et al., 2015]). Target reporters have also been used to investigate whether miRNAs induce correlations in the expression levels of their targets [Bosia et al., 2015]. However, how endogenous miRNA targets simultaneously respond to miRNA induction in individual cells is insufficiently understood, leading to ongoing debates about the influence that one target can have on the expression of others.

In this study we developed a methodology to characterize the regulatory effects of a miRNA on its hundreds of targets in single cells. We constructed an experimental system in which the expression of a miRNA precursor can be induced with doxycycline together with that of GFP from a bidirectional promoter. This system was initially tested with two protein-coding genes, one for the nerve growth factor and the other for eGFP, which showed good, though not perfect correlation at single-cell level [Bornkamm et al., 2005]. In our case, absolute quantification of the miRNA and GFP mRNA in cell populations indicated that expression of the two RNAs was highly correlated in response to doxycycline induction, and we thus

used the GFP mRNA as a proxy for the miRNA. It is likely that a direct measurement of miRNA expression in the cells whose mRNAs are sequenced would further increase the accuracy of the results of our model, and we expect the technology to become available in the near future. We showed that this system exhibits predicted behaviors such as a peak in target noise as well as a peak in the correlation of target levels, in the region of maximal sensitivity to the miRNA. The construct can be easily modified to enable inducible expression of other miRNAs. We further developed a methodology for the variational fitting of Michaelis-Menten-type constants (K_M) characterizing individual miRNA targets. This method takes advantage of the variability in transcriptional activity between individual cells that leads to variability in miRNA expression levels between cells. For the first time we have uncovered the hierarchy of targets of a miRNA, defined by the miRNA concentrations at which these targets respond within the context of all other targets in the cell as well as by the Michaelis-Menten-type constants. We found that high K_M targets undergo the largest down-regulation, indicating that this parameter reflects primarily their k_{cat}, the rate of decay in the presence of the miRNA. Some targets were particularly sensitive to the miRNA, requiring relatively low miRNA concentrations to respond and having reproducibly low A_F^C values. Their higher prediction scores and enrichment in DNA-binding factors suggest that these are prototypical miRNA targets [Gruber and Zavolan, 2013]. Simulations indicate that targets with low K_M and low A_F^C values could sequester the miRNA from other targets if they are highly expressed and do not decay substantially when they interact with the miRNAs. Current approaches for studying miRNA-target interactions, that measure mRNA level changes upon miRNA overexpression to uncover the most relevant targets likely overlook these targets. Thus, it would be interesting to apply our approach to systems in which functional ceRNAs have been reported [Poliseno et al., 2010, Cesana et al., 2011, chun Cheng and Lin, 2013]. Interestingly, early analyses of miRNA and target expression found that many miRNA targets are expressed at relatively high level in the tissue in which the miRNA is expressed [Farh et al., 2005]. However, this has been attributed to miRNAs optimizing the protein output of their targets rather than entirely suppressing it. Our analysis also suggests that targets with low A_F^C, which bind the miRNAs but do not undergo substantial down-regulation in response to it, could impose a threshold for miRNA-dependent regulation, which would otherwise affect a large fraction of the transcriptome.

To demonstrate the robustness of our approach we have inferred parameters of individual targets from two closely related cell lines. However, it is likely that the sensitivity of a target to miRNA regulation is context dependent [Erhard et al., 2014]. Because we wanted to map the parameters of miRNA-target interaction in a native context of mRNA expression, we induced the miRNA expression from an exogenous construct in HEK 293 cells. Although a large number of studies of miRNA-dependent gene regulation have similar designs, it remains possible that

the "true" targets of the miRNA are not naturally expressed in the cell type in which the experiment is carried out. To fully address this possibility one would perhaps have to progressively remove a highly abundant, cell type-specific miRNA, which would be more challenging than inducing miRNA expression. miR-122 in liver cells could be a good candidate for this type of experiment (see also [Denzler et al., 2014]).

The miRNA target parameters that we inferred in our study will enable an improved understanding of the dynamics of networks containing many competing miRNA targets. Furthermore, the approach can be easily extended to RNA-binding protein regulators of mRNA stability as well as to other types of regulators such as transcription factors.

3.4 Methods

3.4.1 A model to describe the dynamics of miRNA targets

We used the model shown in Eqs. (3.1), (3.2) in the main text and also shown below, which considers M targets of a miRNA, each being described by a transcription rate α, decay rate δ, rate of binding the Ago-complexed miRNA k_{on_i}, rate of dissociating from this complex k_{off_i} and rate of degradation when complexed to the miRNA k_{cat_i}. With m_i being the concentration of the free target, A_{m_i} the concentration of the miRNA-bound target, and A the total concentration of Ago-miRNA complexes, we have the following system of $2M$ differential equations

$$\frac{\partial m_i(t)}{\partial t} = \alpha_i - \delta_i m_i(t) - k_{on_i} m_i(t) \left(A - \sum_{j=1}^{M} A_{m_j}(t) \right) + k_{off_i} A_{m_i}(t) \quad (3.4)$$

$$\frac{\partial A_{m_i}(t)}{\partial t} = k_{on_i} m_i(t) \left(A - \sum_{j=1}^{M} A_{m_j}(t) \right) - k_{off_i} A_{m_i}(t) - k_{cat_i} A_{m_i}(t) \quad (3.5)$$

Denoting the total concentration of mRNA i (either free or bound to the miRNA) by T_i and summing the two equations corresponding to mRNA i, the dynamics of T_i is described by

$$\frac{\partial T_i(t)}{\partial t} = \alpha_i - \delta_i m_i(t) - k_{cat} A_{m_i}(t) \quad (3.6)$$

or, in terms of the fraction f_i of molecules of mRNA i that are bound to miRNAs,

$$\frac{\partial T_i(t)}{\partial t} = \alpha_i - \delta_i (1 - f_i) T_i(t) - k_{cat_i} f_i T_i(t). \quad (3.7)$$

Defining the total concentration of mRNA i when no miRNA is present as $T_i^0 = \frac{\alpha_i}{\delta_i}$ and when the miRNA is in high excess as $T_i^\infty = \frac{\alpha_i}{k_{cat_i}}$ we obtain the total concentration of mRNA i at a steady state as

$$T_i^* = \frac{\alpha_i}{\delta_i(1-f_i) + k_{cat_i} f_i} = \frac{T_i^0}{1 + f_i\left(\frac{T_i^0}{T_i^\infty} - 1\right)}. \tag{3.8}$$

Note that the concentration of the miRNA is reflected in the fraction of bound targets. In our experimental system, we vary the expression of the miRNA from very low to very high levels and we can therefore estimate T_i^0 and T_i^∞. However, the fraction of mRNA i that is bound to the miRNA depends not only on the constants of interaction of this mRNA with Ago-miRNA complexes, but also on all other targets that are present in the system. To determine the interaction constants we first derive for each mRNA i the fraction f_i that is bound to the miRNA, as follows. At equilibrium, we have

$$m_i k_{on_i} A_F = A_{m_i}\left(k_{onff_i} k_{cat_i}\right), \tag{3.9}$$

$$\frac{m_i}{A_{m_i}} = \frac{1-f_i}{f_i} = \frac{k_{off_i} + k_{cat_i}}{k_{on_i} A_f}, \tag{3.10}$$

and thus

$$f_i = \frac{1}{1 + \frac{K_{M_i}}{A_F}}, \tag{3.11}$$

with the Michaelis-Menten parameter defined as $K_{M_i} = \frac{k_{off_i} + k_{cat_i}}{k_{on_i}}$. Considering all cells $j \in \{1, ..., N\}$, each with a different concentration of free Ago-miRNA complexes A_F, and substituting f_i in equation (3.8) we obtain

$$T_{ji} = \frac{T_i^0}{1 + \frac{1}{1+\frac{K_{M_i}}{A_{F_j}}}\left(\frac{T_i^0}{T_i^\infty} - 1\right)} \tag{3.12}$$

where T_{ji} is the total concentration of mRNA i in cell j which can be computed from the measured target levels. We isolate the ratio $\frac{K_{M_i}}{A_{F_j}}$ and rewrite

$$\frac{\frac{T_i^0}{T_i^\infty} - 1}{\frac{T_i^0}{T_{ji}} - 1} - 1 = \frac{K_{M_i}}{A_{F_j}} \tag{3.13}$$

or equivalently, in vector form, substituting the left hand side of the equation by $\tilde{T}$,

$$\tilde{T} = A_F^{-1} K_M. \tag{3.14}$$

Here K_M is a $(1 \times M)$-matrix of Michaelis-Menten constants for individual mRNAs, A_F^{-1} is a $(N \times 1)$-matrix of free Ago-miRNA complexes in individual cells and $\tilde{T}$ is a $(N \times M)$-matrix of expression levels of individual mRNAs in individual cells. $\tilde{T}$ can be viewed as a Kronecker product of the two vectors K_M and A_F^{-1}, written in a more general form as

$$\mathbf{B} = xy^\top. \tag{3.15}$$

Determining the vectors x and y becomes the reverse Kronecker product problem and has a known solution satisfying

$$\min_{x,y} \|\mathbf{B} - xy^\top\|_F \tag{3.16}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The solution is obtained from the singular value decomposition (SVD) $\mathbf{B} = U\Sigma V^\top$ as

$$x_i = \sqrt{\Sigma_{11}} U_{i1},\ y_i = \sqrt{\Sigma_{11}} V_{i1} \tag{3.17}$$

From equation (3.13) we see that the SVD provides us the solution (A_F, K_M) up to a scaling factor a, $\frac{aK_M}{aA_F} = \frac{K_M}{A_F}, \forall a \in \mathbb{R}$. In principle, it is possible to determine the factor a which explains the data best, using the total concentration of Ago-miRNA complexes A in all cells.

Fitting the vectors A_F and K_M on data generated from simulations of model (3.4) & (3.5), we found that the correlation of the fitted A_F with the input value was significantly higher than for K_M. This is explained by the fact that we use the total concentration of the miRNA in the cells to sort the cells and smoothen the mRNA expression. A_F being a monotonic, strictly increasing, continuous function of A, smoothing the data along the cell dimension (i.e. along the j index in equation (3.12)) leads to a reduction of noise in the direction of the miRNA levels A_F, but not in the mRNA dimension K_M. Therefore, the vector A_F is inferred more precisely compared to K_M. Using the more precisely inferred A_F values and averaging over cells, we can increase the precision of K_{M_i} values; relation (3.14) always holds and after fitting, we use the values of A_F to compute the values K_M by averaging $A_{F_j}\tilde{T}_{ij}$ over all cells $j \in 1...N$

$$K_{M_i} = \frac{1}{N}\sum_{j=1}^{N} A_{F_j}\tilde{T}_{ji},\ j = 1, ..., N. \tag{3.18}$$

We define $A_{F_j}^C$ the concentration of free Ago at which the target will be exactly halfway between T_i^0 and T_i^∞.

$$\frac{T_i^0}{1 + \frac{1}{1+\frac{K_{M_i}}{A_{F_j}^C}}\left(\frac{T_i^0}{T_i^\infty} - 1\right)} = \frac{T_i^0 + T_i^\infty}{2} \Rightarrow A_{F_j}^C = \frac{K_{M_i}}{\frac{T_i^0}{T_i^\infty}}. \tag{3.19}$$

3.4.2 *In silico* analysis

Stochastic simulations based on equations (3.1) & (3.2) were used to verify the solution obtained in equation (3.12). Stochastic simulations were performed using StochKit v.2.0.11 [Sanft et al., 2011]) with a tau leaping algorithm. For each *in silico* cell, 6 simulations of length 100'000 (arbitrary time units) were carried out to ensure convergence. The first 10'000 steps were considered the "burning phase", and were discarded before the analysis. Means and standard deviations were calculated from the values obtained in the independent simulations.

To test the K_M inference method we constructed an *in silico* data set as follows. We considered a regulatory network of 300 miRNA targets. Each target was characterized by parameters $\alpha_i, \delta_i, k_{cat_i}, k_{on_i}, k_{off_i}$, whose values we assumed to be in the ranges provided by our previous literature survey [Hausser and Zavolan, 2014]. For each target we chose a set of parameters from log-normal distributions, which are shown in Appendix Figure 3.9. Similar to the experimental data set, we considered 4000 virtual cells, each with a distinct concentration of free Ago-miRNA complexes, chosen from a uniform distribution on the $\log_2$ range of -40 to 14, such as ∼50% of cells end up with no miRNA expression, as observed in the experiment. The expression of all targets as a function of the miRNA abundance in these virtual cells is presented in Figure 3.3A. Note however that in the experimental system we could not measure miRNA levels but rather the copy number of the GFP mRNA and thus, in comparing the response of targets in the *in silico* and experimental systems the x-axes differ, being the miRNA level for the *in silico* data, and the GFP mRNA level for the experimental data. Interestingly, the miRNA-to-GFP mRNA conversion factor corresponds well with the miRNA:GFP mRNA ratio of 4-8 that is apparent from the qPCR data (see also Figure 3.1). Each target starts to decay at a specific threshold, depending on its parameters of interaction with the miRNA and the effective miRNA concentration, which depends on the other targets as well. To complete our *in silico* data generation we added log-normal noise to the analytically computed expression levels of the targets (see Figure 3.3C).

To focus on cells were the miRNA targets responded most sensitively to the miRNA, we started with the selection of single cells from which to construct the matrix $\tilde{T}$. T_i^∞ and T_i^0 were calculated from about 200 *in silico* cells with the highest and 1600 cells with the lowest concentration of miRNA, numbers similar to these in the experimental system. We analyzed the derivative of the sum of $\log_2$ target levels in function of miRNA expression and selected the cells where the gradient was lower than -0.01 (Figure 3.3B). Cells with target expression values very close to T_i^0 or T_i^∞ were filtered out to avoid instabilities caused by division by small numbers (see Equation 3.13). Next, we applied a smoothing procedure to ensure that at intermediate miRNA expression, the T_{ji} level of targets i in cells j was strictly in the range $(T_i^\infty; T_i^0)$ (see Figure 3.3C). We started by replacing the

expression level of a given target in a given cell with the mean over the 50 cells with miRNA expression level closest to that in the reference cell. In a second pass, for the smoothed T_{ji} values outside of the $(T_i^\infty; T_i^0)$ range, we computed again a running mean starting with a window size of ten and discarding iteratively the strongest outliers until the mean value T_{ji} within each window was within the $(T_i^\infty; T_i^0)$ range. For the windows were this procedure did not leave any points, we increased the size of the second-pass window locally, repeating the pruning procedure until all the T_{ji} values were within the $(T_i^\infty; T_i^0)$ range. To ensure the stability of the SVD we adjusted the boundary of the T intervals computed from the data by a small safety margin c (*i.e.* $T_i^0 - c < T_{ji} > T_i^\infty + c, c = 10\%$ of $T_i^0 - T_i^\infty$ for each gene).

We assessed the accuracy of the fitting procedure by comparing the inferred $A_{F_j}^C$ and K_{M_i} parameters with those that were used in the model that generated the *in silico* data. In spite of very high noise (Figure 3.3C) there was a good correlation between the fitted and input values of the parameters, as shown in Figure 3.3D. In addition, the correlation of parameters observed when simulating two independent "samples", with two independent noise applications, was also relatively high (Figure 3.3D). We also observed that the range of inferred K_Ms is narrower than the range of input K_Ms.

Having validated our inference procedure on *in silico* data, we applied it to the experimental data.

3.4.3 Cell culture

We used a Human Epithelial Kidney (HEK) 293 cell line with inducible expression of hsa-miR199a (i199) and a derivative of this cell line (i199KTN1) in which a Renilla Luciferase coding sequence followed by the 3'UTR of the kinectin 1 gene (KTN1) were inserted in the genome. These cell lines have been introduced in a previous study [Hausser et al., 2013]. Cells were grown in DMEM media with 10% FCS supplemented with Pen-Strep and Hygromycin for plasmid integrity. For all the experiments, unless otherwise mentioned, cells were stimulated with doxycycline at concentrations of 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003 or 0 $\mu g/ml$, for 8 consecutive days. During this period, fresh medium with doxycycline was provided every 24 hours and cells were split every 72 hours to prevent slowed down growth in confluent cultures [Ghosh et al., 2015].

3.4.4 Single cell mRNA-sequencing

Cell capture, GEM Barcoding and cDNA synthesis

Cells were detached with Accutase® reagent (Gibco, Life Technologies™). The cell number was determined with the Countess™ Automated Cell Counter (invitrogen™)

following manufacturer's instructions. Cells that were induced with different doxycycline conditions (see section above) were pooled together in equal proportions (1500 cells/μl of each). The cells were finally resuspended in PBS containing 0.04% BSA at a target concentration of 700 cells/μl after straining with a cell strainer to avoid clumps. This is performed so as to partition the input cells across tens of thousands of droplets (GEMs) for the purpose of lysis and barcoding. GEM Generation & Barcoding was performed according to manufacturer's instructions (Chromium^{TM} Single Cell 3' Reagent Kits v2, Part No-120234, 10X Genomics). Subsequently, reverse transcription (RT), and post GEM-RT cleanup was done exactly as specified in the protocol from the manufacturer. The purified GEM-RT product was then pre-amplified for 10 cycles, purified with SPRI select (Beckman Coulter) and analysed on a high sensitivity Bioanalyzer.

Library preparation and sequencing

Library construction including Fragmentation, End Repair & A-tailing was performed as per manufacturer's protocol (Chromium^{TM} Single Cell 3' Reagent Kits v2, Part No-120234, 10X Genomics). Subsequently the fragments were purified with a double sided size selection with SPRI select (Beckman Coulter), and ligated to adapters. After ligation, the samples were purified once more with SPRI select prior to the steps of sample index PCR reactions. The end product was finally obtained with another round of double sided SPRI selection of the PCR reaction. Quality control of the libraries was done on an Agilent Bioanalyzer High Sensitivity chip. Libraries were then sequenced (Paired End) on a NextSeq 500 system (NextSeq 500/550 High Output v2 kit (75 cycles)) and the reads were obtained according to the following parameters :

1. Seq Read1 26 cycles
2. Seq Read2 58 cycles
3. IDX Read 8 cycles
4. Illumina basecalling software version: bcl2fastq v2.19.0.316
5. Demultiplexing software version: cellranger mkfastq (2.0.0)

The library preparation and sequencing were performed at the Genomics Facility Basel. The sequencing data has been deposited to the Sequence read archive (`www.ncbi.nlm.nih.gov/sra/`) under the accession number SRP067502.

Computation of the coefficient of variation of target expression

Given the set of cells sorted by their GFP expression, we calculated the coefficient of variation (CV, standard deviation/mean) of a specific target as follows.

We traversed the list of cells from those with lowest to those with highest GFP expression and for each cell, we considered the 199 cells with closest GFP level to the reference cell, and calculated the CV of each target. We then $\log_2$-transformed the CV of individual targets and determined the mean (and standard error) over all 100 selected low A_F^C targets. We applied the same procedure to all non-targets (genes targeted neither by hsa-miR-199-3p nor by hsa-miR-199-5p). We then subtracted the $\log_2$ mean CV of targets and non-targets, repeated this procedure for the entire GFP expression range and shown the normalized CV as a function of the $\log_2$ GFP level in the reference cell.

PAGODA variance normalization

The 199 and i199-KTN1 single cell data sets were divided in 100 cell batches, grouped according to GFP expression level in the cells. A random sample of 100 cells was subsampled from the cell population with no GFP expression. Next, the PAGODA data preparation, error modeling and variance normalization functions were used with a standard parameters, on the raw data sets, as specified in the PAGODA tutorial, `http://hms-dbmi.github.io/scde/pagoda.html`. The normalized variance was used for the analysis showed in Appendix Figure 3.8B,E.

Computation of the pairwise correlation coefficients of target expression levels

Given a population of cells sorted by their GFP expression, we calculated the Pearson correlation of $\log_2$ expression levels for all pairs of 100 targets, in function of GFP level (as for CV, average values were computed over 199 cells with GFP expression closest to that in the reference cell). Thus, we started from those cells with lowest GFP expression and moved by one cell at a time to cells with the highest GFP expression, computing the mean correlation coefficient (and standard error of the mean) over all pairs of genes within a cell. We repeated the procedure for 50 evaluations of 100 random genes that were not predicted as targets. Finally, we divided the mean correlation coefficients of targets and non-targets and shown this as function of GFP level in the cell.

Computation of GO enrichment

The hyperGTest function from GOstats package (R-Bioconductor repository) was used to find enriched GO terms. The maximum 'pvalueCutoff' for reporting was set to 0.1, 'conditional' to 'TRUE' and 'testDirection' to 'Over'.

3.4.5 Cell population mRNA-seq

Total RNA isolation

Total RNA was extracted with TRI Reagent® (Sigma-aldrich) following manufacturer's instructions. Briefly, cells were detached from the plate by 5 min incubation with Trypsin-EDTA solution (T3924 SIGMA), conditioned media was added and whenever necessary, cells were counted with a Countess™ cell counter (Thermo Fisher Scientific). A defined number of cells were pelleted and either snap frozen for future use or resuspended right away in TRI Reagent® (#T9424,Sigma-aldrich). Total RNA was resuspended in nuclease free Water (#AM9937, Thermo Fisher Scientific). Samples were always kept on ice or at -80°C.

mRNA purification

To select the Poly(A)$^+$ RNA, a double purification with Dynabeads® Oligo (dT)25 (Dynabeads® mRNA DIRECT™ Kit, Ambion™) was performed, using the manufacturer's manual and recommendations. Since the starting material was purified total-RNA, only buffer B was used for the washing steps.

Library preparation

Purified mRNA was fractionated with Alkaline Hydrolysis Buffer at 95°C for 5 min. Fractionated mRNA was selected with RNeasy MinElute Cleanup Kit (Qiagen, Inc.). Purified mRNA fragments were dephosphorylated with FastAP (Life Technologies, Inc.) and 5'-phosphorylated with PNK (Life Technologies, Inc.) following manufacturer's instructions for optimal conditions of the enzymatic reaction. After another round of RNeasy MinElute Cleanup Kit (Qiagen, Inc.), a pre-adenylated DNA adapter (5'-TGGAATTCTCGGGTGCCAAGG-3') was ligated to the 3'end of the mRNA fragments at 4°C overnight using the T4 RNA ligase 2, truncated K227Q (New England Biolabs, Inc.), in 1× T4 RNA ligase buffer (no ATP) and 15% DMSO. The next day, after another round of RNeasy MinElute Cleanup Kit (Qiagen, Inc.), an RNA adapter (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3') was ligated to the 5'end of the RNA fragments at 4°C overnight using the T4 RNA ligase 1 (Life Technologies, Inc.), in 1× T4 RNA ligase buffer ($1mM$ ATP) and 15% DMSO. Next day, after another round of RNeasy MinElute Cleanup Kit (Qiagen, Inc.), Reverse Transcription was performed using Superscript IV (Invitrogen, Inc.) and RTP primer (5'-CCTTGGCACCCGAGAATTCCA-3'), following manufacturer's instructions. cDNA was then amplified by 12 cycles of PCR using NEBNext® High-Fidelity 2× PCR Master Mix (NEB, Inc.), and Illumina TruSeq® Small RNA PCR compatible primers.

Library sequencing

The library was sequenced in the Genomics Facility Basel, on Illumina HiSeq 2000 or HiSeq 2500 instruments using Truseq compatible primers. Reads of 50 nt were generated along with 8nt index reads corresponding to the sample-specific barcode.

3.4.6 Read mapping and data preprocessing

Reads from single cell and cell population mRNA-Seq were mapped to the transcriptome (Ensembl, GRCh38.rel88) with Cellranger-1.3.1, the software provided by 10x Genomics to map the reads produced by the Chromium™ Single Cell 3' solution. Cellranger processes the cell and transcript barcodes, uses STAR 2.5.1b to align the reads and counts the number of transcripts observed from each gene to provide a table of unique molecular identifier (UMI) counts per gene and per cell. The sequence of the eGFP mRNA, that was expressed from the exogenous pRTS1 vector, was added to the transcriptome before mapping. After summing the counts for all Ensembl entries for a given Entrez gene ID, the gene counts were normalized to have in each cell one million counts. Next, a pseudocount, 0.001, was added to each gene (and 1.0 to GFP gene for clarity of visualization). In all of the analyses, genes with very low final estimated expression (mean TPM < 7 across cells) were discarded.

Targets selection

If not specified otherwise, we used in analyses the 300 highest probability targets predicted by MIRZA-G-C [Gumienny and Zavolan, 2015] that were down-regulated at least 8% at the maximum miRNA concentration ($\log_2(T_i^\infty/T_i^0) < -0.12$). This selection applied to both miRNAs and both cell lines.

3.4.7 mRNA and miRNA qPCR

Cells were induced with various concentrations of doxycycline (as indicated in the figure) for 8 days. After counting the cells, total RNA was extracted with TRI Reagent® (Sigma-aldrich) following manufacturer's instructions. cDNA of the targets of interest was generated using superscript III (Invitrogen™) following manufacturer's protocol. For miRNA assays, reverse transcription and PCR of either non-induced or Dox induced cells were performed following the TaqMan® Small RNA Assays quick reference protocol (Life Technologies™) with 100 *ng* of total RNA. For estimation of relative miRNA quantities , hsa-miR-16 levels were used as an invariant control. For reverse transcription of GFP mRNA, the following linear DNA primer was used: EGFP_R RT taqman, 5'-TGTCGCCCTCGAACTTCAC-3'. To generate a cDNA copy of hsa-miR199a-5p a stem-loop primer system from Life technologies ™ was used (Assay ID-000498).

All qPCR were performed and read in StepOnePlus™ Real-Time PCR Systems (Life Technologies™). To obtain absolute quantification data, standard curves for GFP and hsa-miR-199a-5p were also included. GFP mRNA was generated by *in vitro* transcription with pcDNA3-eGFP linearized vector and RiboMAX™ Large Scale RNA Production System – T7 (Promega, Co.) following manufacturer's instructions. Molarity was estimated taking into account mass concentration (Qubit® RNA HS assay kit - Life Technologies™), average length (Agilent RNA 6000 Pico Kit - Agilent Technologies, Inc.) and fragment sequence, with the following formula: molarity = mass/(length×mass RNA base). The hsa-miR-199a-5p miRNA (5'-CCCAGUGUUCAGACUACCUGUUC-3') was ordered from Microsynth AG, and the molarity was calculated the same way. Absolute molecule numbers were obtained utilizing the StepOne™ Software (Life Technologies™).

3.4.8 CLIP Seq

CLIP Seq was performed as described in Jaskiewicz L et. al [Jaskiewicz et al., 2012] with few modifications. Ago2-CLIP in i199 cells was performed using Ago2 antibody-containing serum (kind gift from Prof Gunter Meister, University of Regensburg, Germany) crosslinked to 100 μl of Dynabeads Protein G (#10003D, Thermo Fisher Scientific). TURBO DNase (#AM2238, Thermo Fisher Scientific) treatment of UV-crosslinked cell lysates was followed by a brief treatment with RNase T1 (#EN0541, Thermo Fisher Scientific) for the specific recovery of Ago2-protected RNA fragments. Subsequently, antibody-bound beads were incubated with the cell lysate for 3 hrs at 4°C for precipitation. Furthermore, the beads were washed, treated again with RNase T1, dephosphorylated and labelled with radioactive ATP [γ-32P] to facilitate purification of the required fragments from a nitrocellulose membrane filter following a standard SDS PAGE electroelution process. The recovered RNA fragments were ligated to a pre-adenylated 3' adapter, annealed to the RT primer and subsequently ligated to the 5' adapter prior to a step of reverse transcription with SuperScript™ III Reverse Transcriptase (#18080044, Thermo Fisher Scientific). In the finals steps, a PCR amplification of the reverse-transcribed cDNA derived from the Ago2 immunoprecipitate, was followed by size selection of 140-180 nucleotide long fragments in native PAGE and sequenced after purification.

Data availability

The datasets produced in this study are available in the following databases:

- RNA-Seq, scRNA-Seq and CLIP-Seq data: NCBI Sequence Read Archive with accession SRP150046 (`https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP150046`)

Acknowledgements

We are grateful to Andrea Riba, Alexander Kanitz, Joao Guimaraes, Andreas R. Gruber and the other members of the Zavolan group for providing input and feedback during the project and for the careful reading of the manuscript. The work was supported by the European Research Council Starting grant 310510-WHYMIR and by the SystemsX.ch systems biology initiative in Switzerland through the RTD grants 51RT-0_145680 (StoNets) and 51RT-0_145728 (NeuroStemX).

Author contributions

Andrzej J. Rzepiela co-developed the mathematical model, carried out the simulations and computational analysis of the experimental data; Souvik Ghosh generated the single cell sequencing data and carried out experiments to validate miRNA expression; Jeremie Breda developed the mathematical model; Arnau Vina-Vilaseca helped set up the experimental system and carried out the CLIP experiments; Afzal P. Syed developed the i199 and i199-KTN1 cell lines and helped with the CLIP experiments; Andreas J. Gruber analyzed the downregulation of miRNA targets in bulk populations; Katja Eschbach and Christian Beisel provided technical help with the single cell experiments; Erik van Nimwegen contributed to the mathematical model; Mihaela Zavolan designed the study; contributed to data analysis and supervised the work. AJR and MZ wrote the manuscript with help from all authors.

Disclosure declaration

The authors declare that they have no conflict of interest.

3.5 Appendix

3.5.1 Supplemental Figures

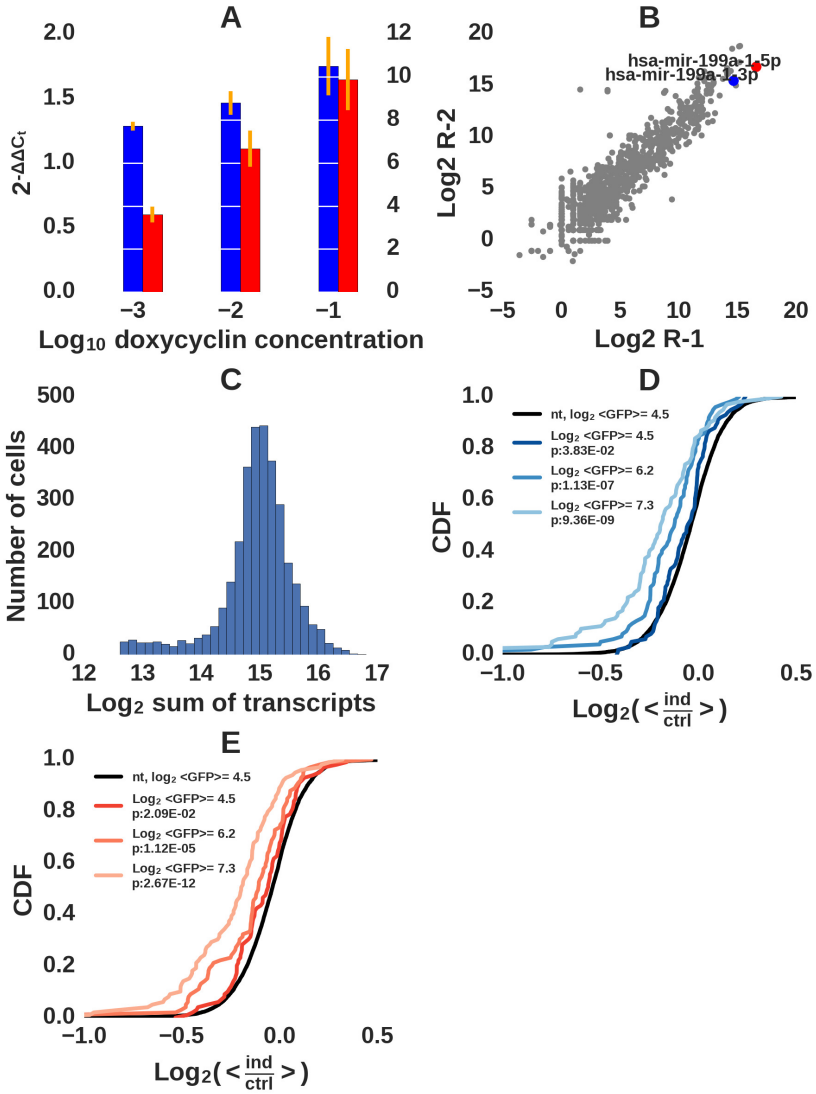


Figure 3.6: **Appendix Figure S1. Characterization of hsa-miR-199a-5p and hsa-miR-199a-3p miRNA activity. A.** Relative hsa-miR-199-3p (blue, left y-axis) and hsa-miR-199-5p (red, right y-axis) miRNA levels in doxycycline-induced cells compared to the non-induced cells, measured by quantitative PCR, demonstrate that the two miRNAs are co-expressed. The C_t values obtained for each set were normalized to the levels of hsa-miR-16 and to the values from non-induced cells. Error bars indicate standard deviations from 2 experiments. **B.** Expression of miRNAs in fully induced HEK cells as measured by Clip-Seq. Two replicates are correlated, hsa-miR-199-5p and hsa-miR-199-3p are indicated. **C.** Number of transcripts identified in individual i199 cells. **D,E.** Downregulation of top 100 predicted targets of the miRNAs in i199 cells with different levels of GFP. Three sets of cells (200 cells each) with increasing GFP expression levels were used, showing that the downregulation of hsa-miR-199-3p (blue lines, **D**), and hsa-miR-199-5p (red lines, **E**) targets increases with level of GFP expression; the distribution of log- fold changes of non targets is shown in black. P values are from the Kolmogorov-Smirnov test comparing the distributions of targets with that of non-targets.

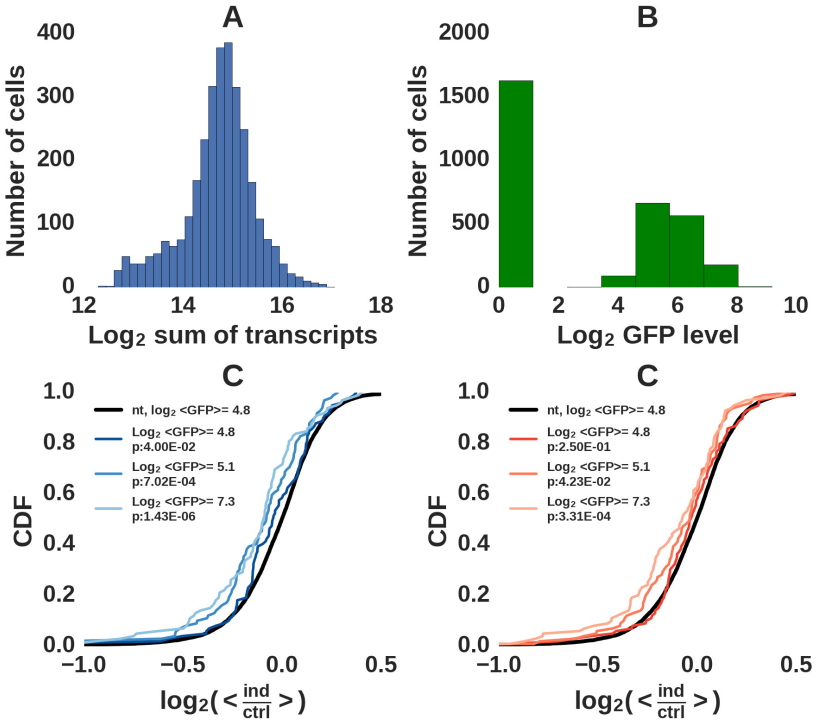


Figure 3.7: **Appendix Figure S2. Characterization of miRNA activity in single i199-KTN1 HEK cells. A.** Count of transcripts identified from each individual gene in single i199-KTN1 cells. **B.** Normalized GFP mRNA expression distribution in single cells. **C,D.** Downregulation of top 100 predicted targets of the miRNAs in i199-KTN1 cells with different levels of GFP. Three sets of cells (200 cells each) with increasing GFP expression levels were used, showing that the downregulation of hsa-miR-199-3p (blue lines, **C**), and hsa-miR-199-5p (red lines, **D**) targets increases with level of GFP expression; the distribution of log-fold changes of non targets is shown in black.

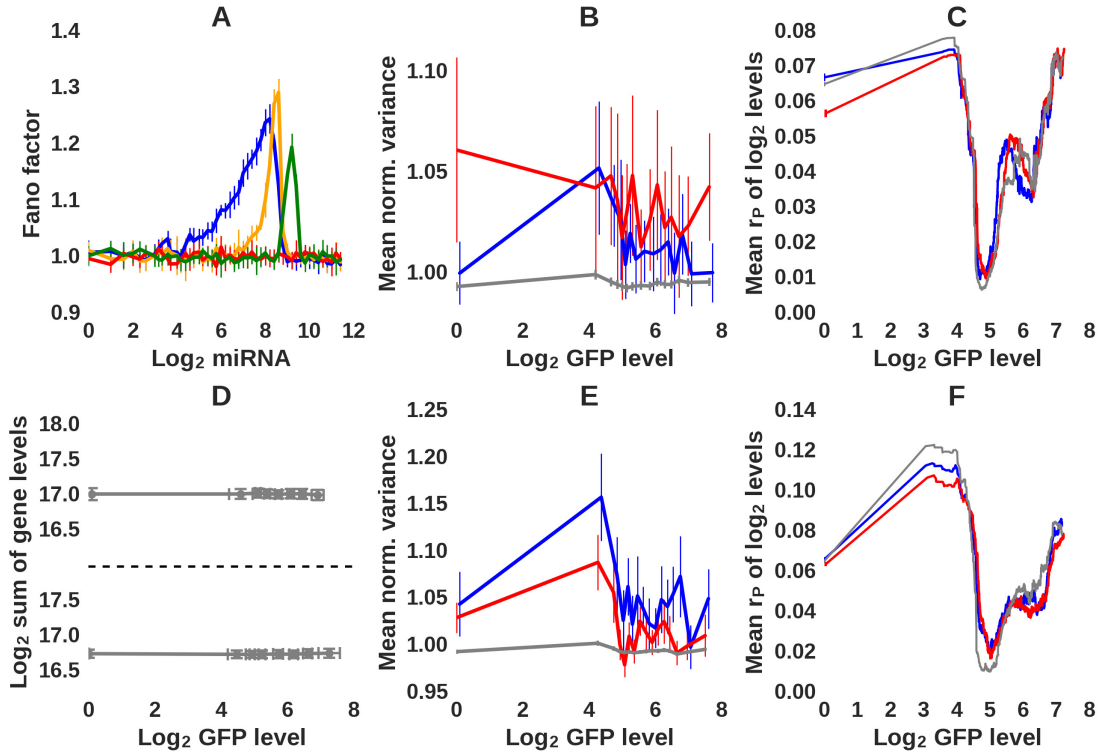


Figure 3.8: **Appendix Figure S3. Expected and observed response of miRNA targets to miRNA induction in single cells; additional information. A.** Fano factor, of *in silico* target levels across cells, calculated in function of miRNA expression, from the simulation trajectories. The panel corresponds to panel B, Figure 3.2, where C_V is calculated using the same data. **B, E.** Normalized variance (using PAGODA package [Fan et al., 2016]) of 100 lowest A_F^C hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets and all genes (grey) in the i199 (**B**) and i199-KTN1 (**E**) cells, in function of $\log_2$ GFP expression in the same cells; see Methods about PAGODA normalization and calculation details. **D.** Total expression ($\log_2$ sum of TPMs) of 1000 random genes in the i199 (lower line) and i199-KTN1 (upper line) cells. **C,F.** Mean Pearson pairwise correlation coefficients for miRNA targets in function of GFP expression in i199 (**C**) and i199-KTN1 (**F**) cells. Mean from 50 calculation evaluations of random selection of 100 non-target genes is shown as grey line. Means were calculated from the two hundred cells with GFP expression closest to a specific expression level (**C,D,F**). For **A,C,D** and **F** panels standard deviations are shown, for **B** and **E** standard error.

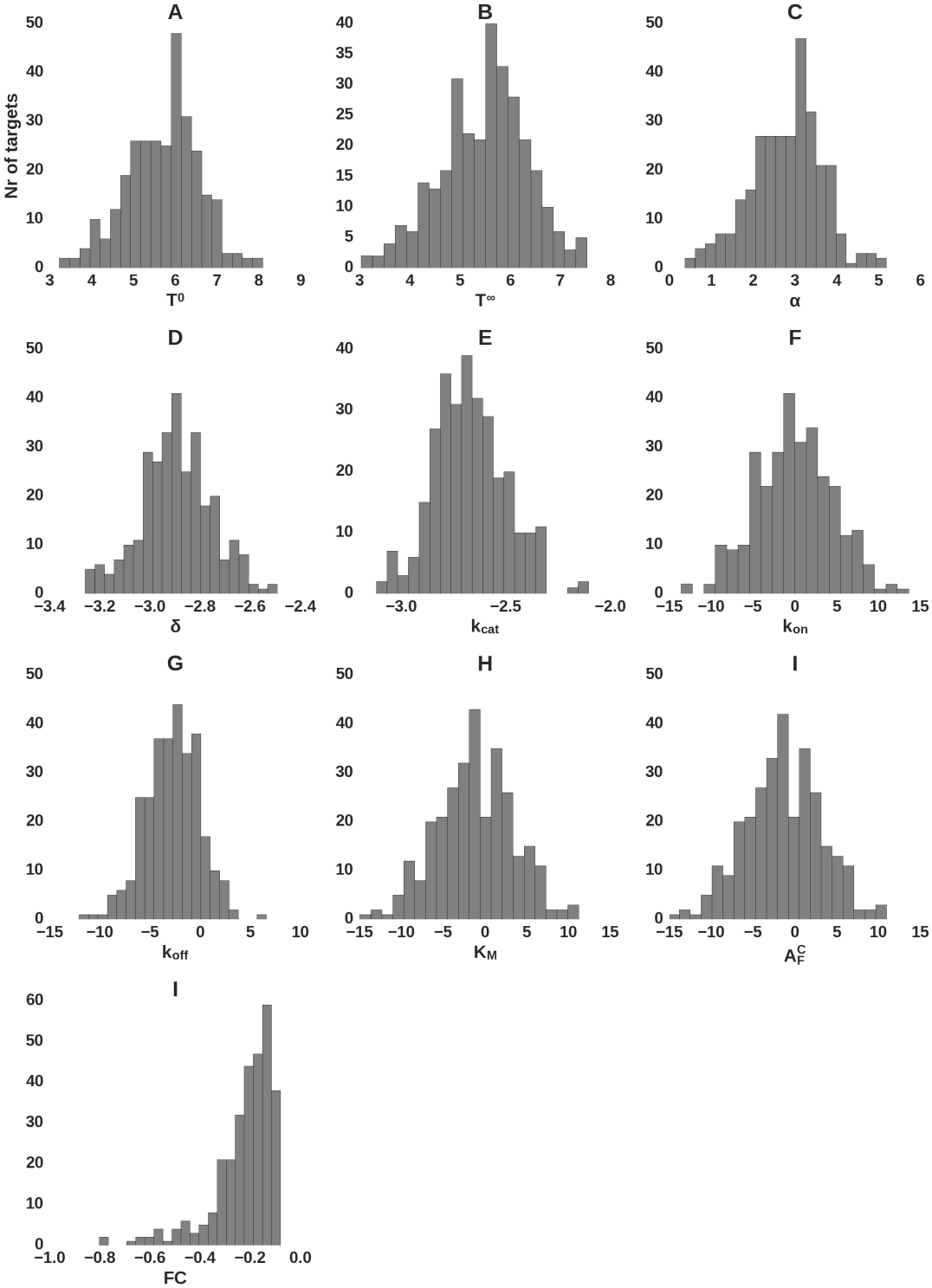


Figure 3.9: **Appendix Figure S4. Distribution of parameters of the *in silico* targets.** $\log_2$ values are shown. See section "*In silico* analysis" for additional explanation.

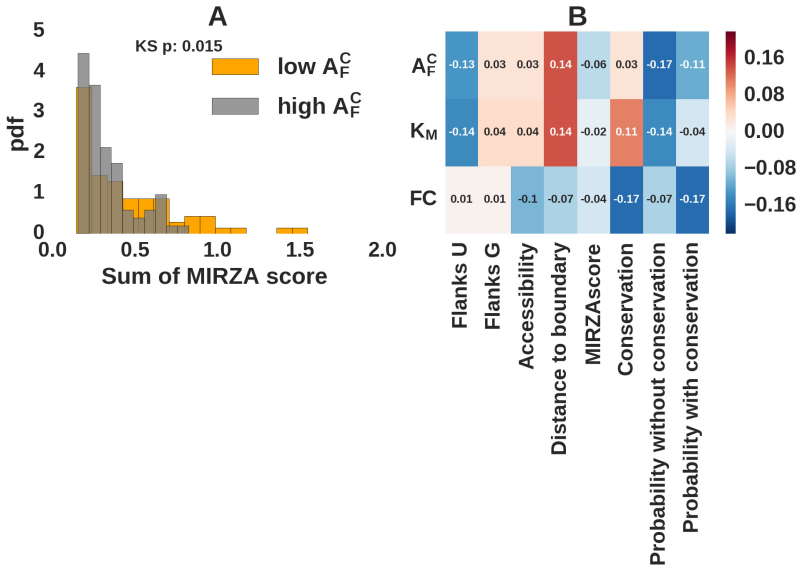


Figure 3.10: **Appendix Figure S5. hsa-miR-199a-3/5p targets A_F^C correlate with binding site properties. A.** Low A_F^C targets have higher MIRZA-G-C scores compared to high A_F^C targets. For each gene, we calculated the mean of the A_F^C values inferred from the i199 and i199-KTN1 data. We also calculated total MIRZA-G-C scores for each gene by summing the prediction scores for the two miRNAs. We took the union of the 40 targets with with lowest A_F^C for the two miRNAs (77 targets, as some were targeted by both miRNAs) and the similar list of targets with the highest A_F^C and compared their MIRZA-G-C scores. **B.** Spearman correlation of A_F^C, K_M and FC to site properties. Binding site properties used to calculate MIRZA-G-C score for all targets of the two miRNAs which have only one binding site (of either of the two miRNAs, about 70% of targets) are used in this correlation. The partial properties are not additive and thus multi-site targets are omitted. Here "Probability with conservation" is the MIRZA-G-C score.

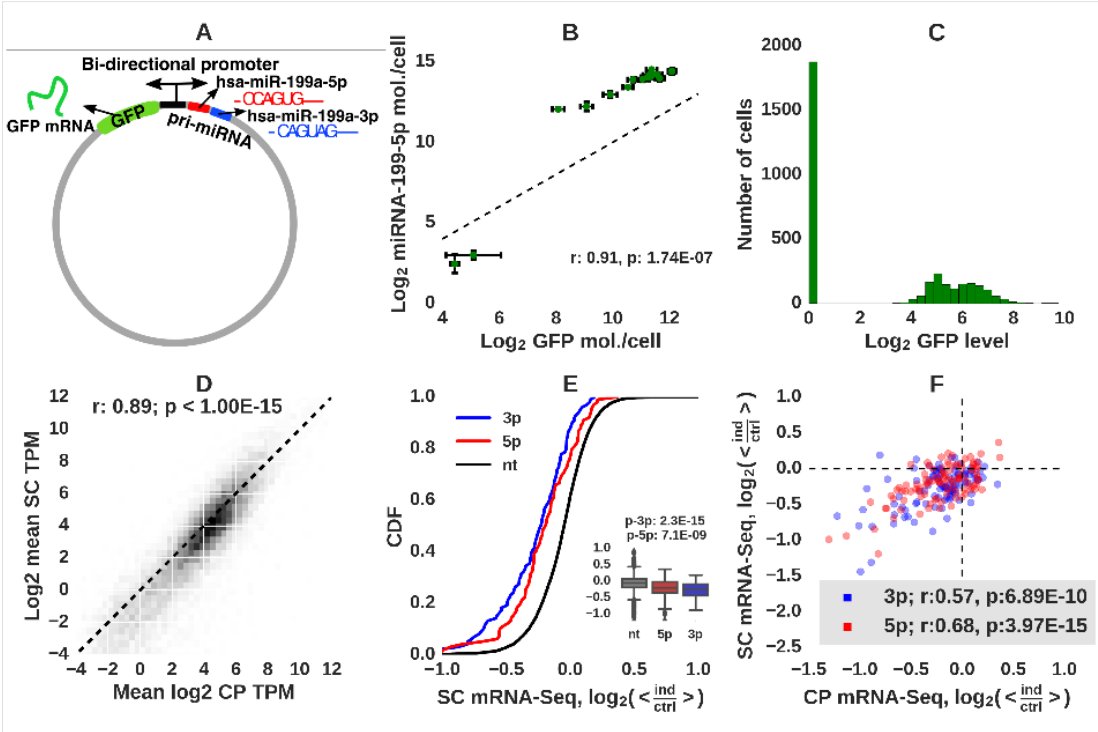


Figure 3.11: **Appendix Figure S6. Design and characterization of the experimental system. A.** Schematic representation of the construct used to express hsa-miR-199a-5p (red), hsa-miR-199a-3p (blue), and the reporter GFP mRNA from a bidirectional promoter. Shown are also the "seed" sequences (nucleotides 2-7) of the two miRNAs. **B.** The expression levels of hsa-miR-199a-5p and GFP mRNA, measured from cell populations by quantitative PCR, are highly correlated. **C.** Histogram of normalized GFP mRNA expression (TPM) in individual i199 cells. **D.** Correlation of mRNA expression levels estimated from SC sequencing (1875 T^0 cells (see text for definition) from which no GFP mRNA was captured) and from CP mRNA-seq (6 replicates of non-induced cell populations). **E.** Cumulative distribution of expression differences of the top 100 targets of hsa-miR-199a-5p (red), top 100 targets of hsa-miR-199a-3p (blue), and of 6179 remaining, "background" genes (black) between cells expressing highest and lowest GFP levels (216 T^∞ cells with > 6.8 TPM GFP ("ind") vs. 1875 T^0 cells with 0 TMP GFP ("ctrl")). Box plots of $\log_2$-fold change of non-targets, top 100 miR-199a-3p and top 100 miRNA-199a-5p targets are shown in the inset. P-values of the rank-sum test comparing targets and non-targets are also shown. **F.** Scatter plot of expression differences of the top 100 targets of each miRNA, estimated from bulk sequencing (CP) or from single cell sequencing (T^∞ and T^0 cells defined as for previous panel). Similar to main Figure 3.1, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

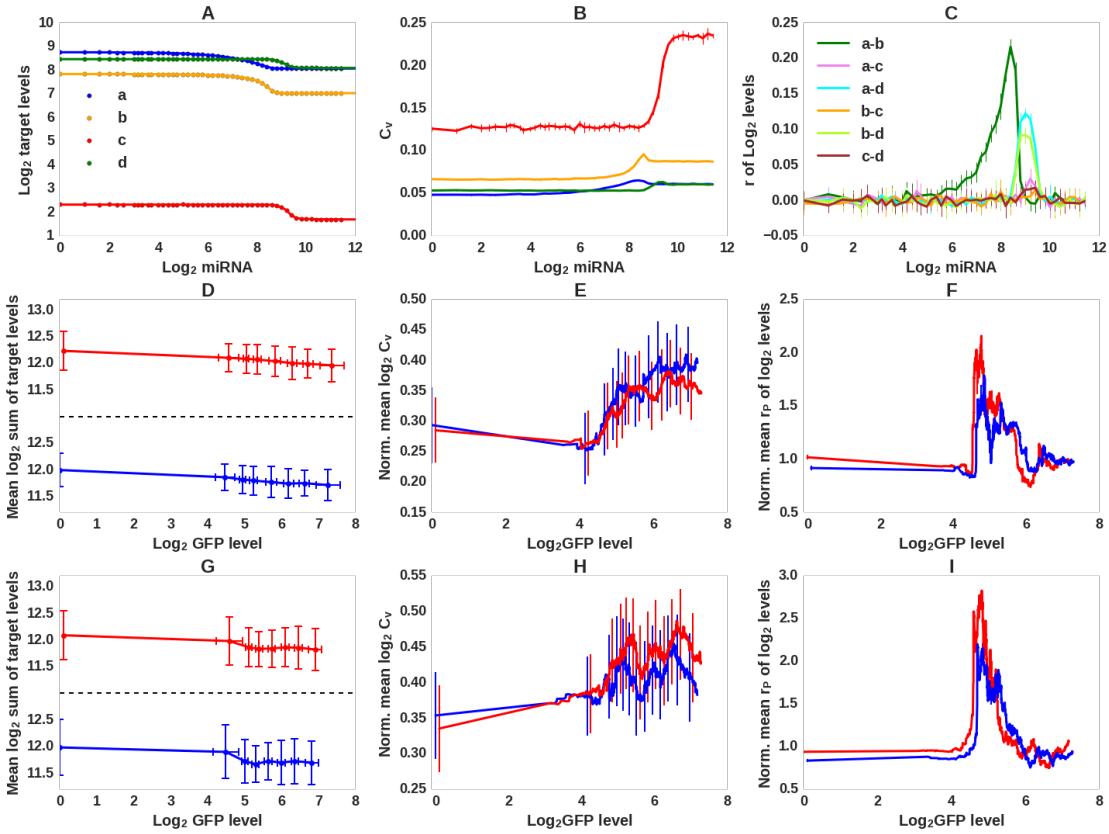


Figure 3.12: **Appendix Figure S7. Expected and observed response of miRNA targets to miRNA induction in single cells. A.** Results of numerical integration (Eqs 3.1-3.2, solid lines) and the average of six stochastic simulations (dots) of a model with four target genes (indicated by distinct colors) chosen to cover a wide expression range and to have either high or low sensitivity to the miRNA. Fifty *in silico* cells, each with a defined miRNA concentration were simulated. **B.** Coefficient of variation (C_V) of *in silico* target levels across cells, calculated in function of miRNA expression, from the simulation trajectories. **C.** Pearson correlation coefficients of expression levels of pairs of genes from *in silico* cells, calculated in function of miRNA expression from the simulation trajectories. **D,G.** Total expression ($\log_2$ sum of TPMs) of 100 lowest A_F^C hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets (see also Methods for target selection) in the i199 (**D**) and i199-KTN1 (**G**) cells, in function of $\log_2$ GFP expression in the same cells. **E,H.** Mean C_V and **F,I.** Mean Pearson pairwise correlation coefficients for miRNA targets in function of GFP expression in i199 (E,F) and i199-KTN1 (**H,I**) cells. Averages were calculated from the two hundred cells with GFP expression closest to a specific expression level. C_V values are shown as ratios to corresponding values computed for all mRNAs (**E,H**) and r_P to mean of 50 evaluations of random selection of 100 control genes (**F,I**). For **B,C,D** and **G** plot standard deviations are shown, for **E,F,H** and **I** standard error. Similar to main Figure 3.2, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

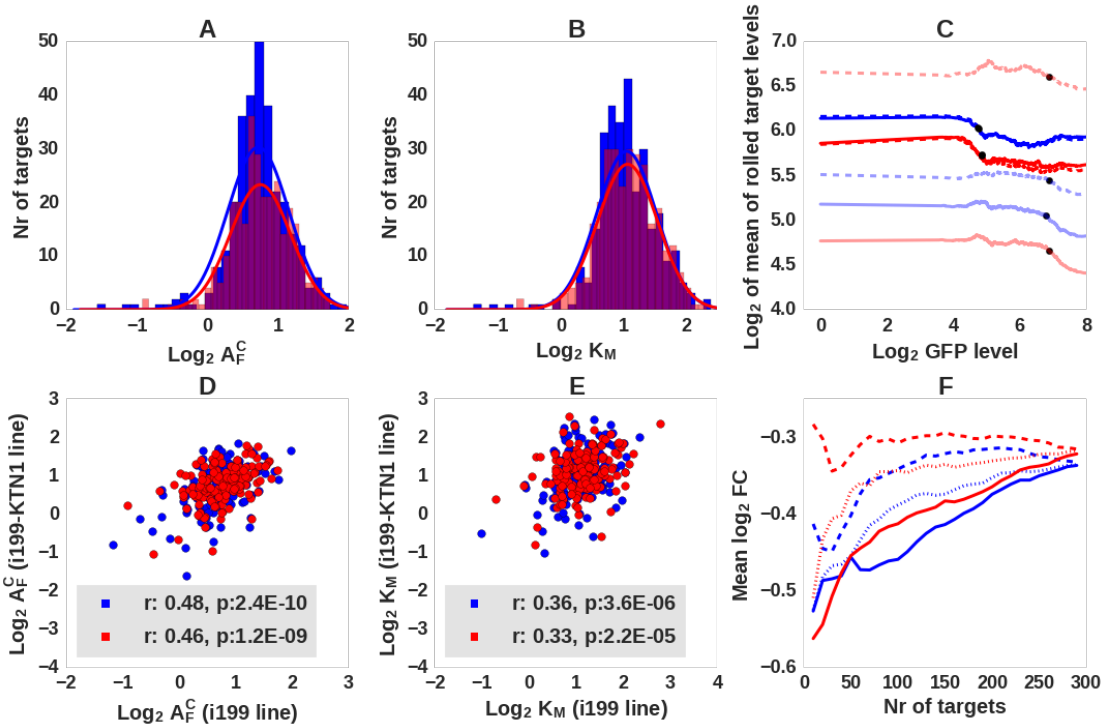


Figure 3.13: **Appendix Figure S8. Parameters describing the response of individual targets to changes in miRNA expression.** Histograms of A_F^C (**A**) and K_M (**B**) values of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets, inferred from the i199 cell line. The lines indicate the best-fitting Gaussian distributions. **C.** Response of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets to the miRNAs in i199 cells. Targets were selected based on A_F^C (dashed lines) or K_M (full lines) values, targets with low values of the respective parameters are shown in strong color and those with high values in faded colors. 20 targets were summed up for each category. Dots show the point where the targets have undergone 1/2 of their maximal down-regulation. **D.** Scatter plot of $\log_2 A_F^C$ values inferred for individual targets from the i199 and i199-KTN1 cell lines. Shown are also Pearson correlation coefficients and corresponding p-values. **E.** Scatter plot of $\log_2 K_M$ values inferred for individual targets from the i199 and i199-KTN1 cell lines. Shown are also Pearson correlation coefficients and corresponding p-values. **F.** Average $\log_2$ fold change of hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets as a function of the number of top targets considered, where predictions are made based either on K_M values (highest to lowest, full lines), A_F^C values (lowest to highest, dashed lines) or Target-Scan context+scores scores (lowest to highest, dotted lines). Similar to main Figure 3.4, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

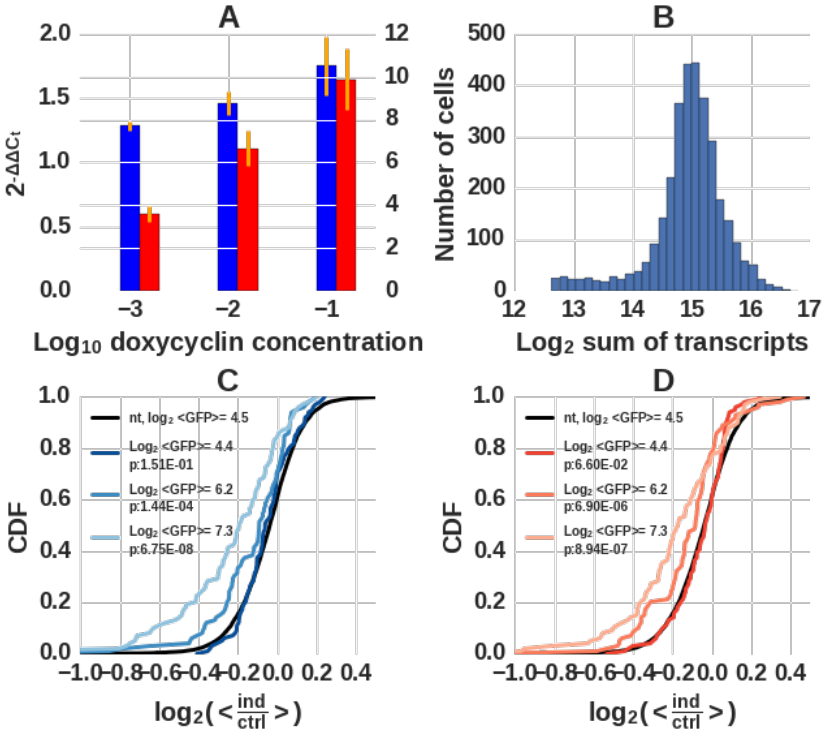


Figure 3.14: **Appendix Figure S9. Characterization of hsa-miR-199a-5p and hsa-miR-199a-3p miRNA activity. A.** Relative hsa-miR-199-3p (blue, left y-axis) and hsa-miR-199-5p (red, right y-axis) miRNA levels in doxycycline-induced cells compared to the non-induced cells, measured by quantitative PCR, demonstrate that the two miRNAs are co-expressed. The C_t values obtained for each set were normalized to the levels of hsa-miR-16 and to the values from non-induced cells. Error bars indicate standard deviations from 2 experiments. **B.** Expression of miRNAs in fully induced HEK cells as measured by Clip-Seq. Two replicates are correlated, hsa-miR-199-5p and hsa-miR-199-3p are indicated. **C.** Number of transcripts identified in individual i199 cells. **D,E.** Downregulation of top 100 predicted targets of the miRNAs in i199 cells with different levels of GFP. Three sets of cells (200 cells each) with increasing GFP expression levels were used, showing that the downregulation of hsa-miR-199-3p (blue lines, **D**), and hsa-miR-199-5p (red lines, **E**) targets increases with level of GFP expression; the distribution of log- fold changes of non targets is shown in black. *P*-values are from the Kolmogorov-Smirnov test comparing the distributions of targets with that of non-targets. Similar to Figure 3.6, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

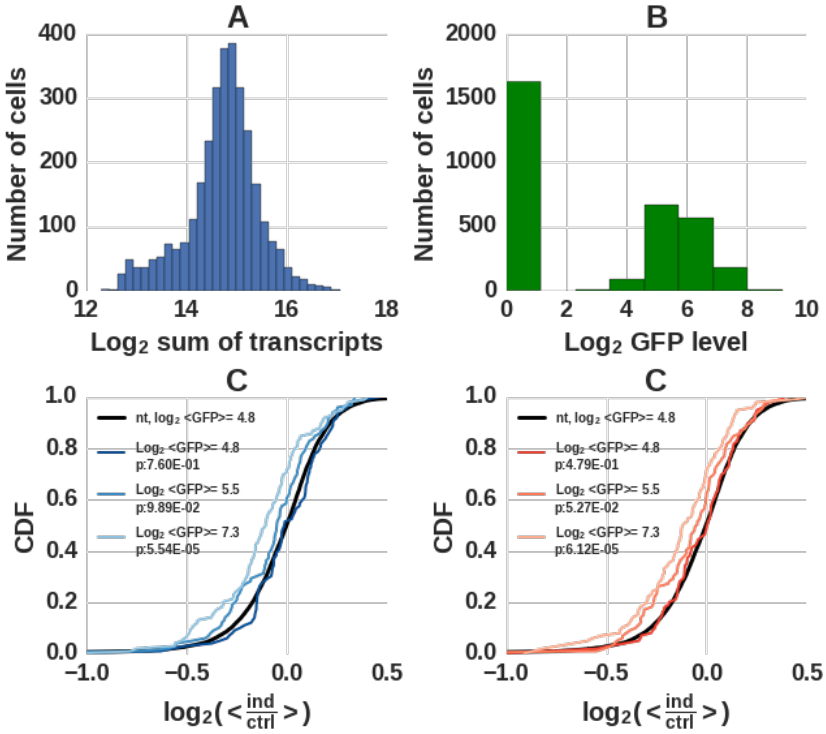


Figure 3.15: **Appendix Figure S10. Characterization of miRNA activity in single i199-KTN1 HEK cells. A.** Count of transcripts identified from each individual gene in single i199-KTN1 cells. **B.** Normalized GFP mRNA expression distribution in single cells. **C,D.** Downregulation of top 100 predicted targets of the miRNAs in i199-KTN1 cells with different levels of GFP. Three sets of cells (200 cells each) with increasing GFP expression levels were used, showing that the downregulation of hsa-miR-199-3p (blue lines, **C**), and hsa-miR-199-5p (red lines, **D**) targets increases with level of GFP expression; the distribution of log-fold changes of non targets is shown in black. Similar to Figure 3.7, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

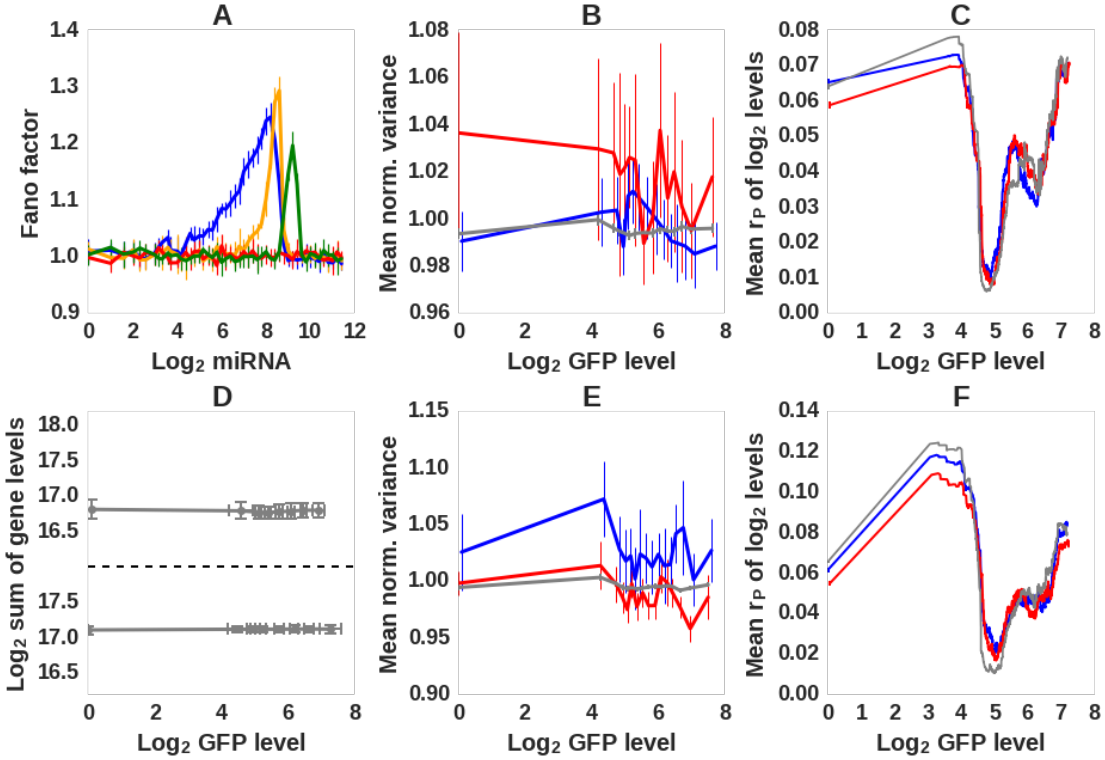


Figure 3.16: **Appendix Figure S11. Expected and observed response of miRNA targets to miRNA induction in single cells; additional information. A.** Fano factor of *in silico* target levels across cells, calculated in function of miRNA expression, from the simulation trajectories. The panel corresponds to panel B, Figure 3.2, where C_V is calculated using the same data. **B,E.** Normalized variance (using PAGODA package [Fan et al., 2016]) of 100 lowest A_F^C hsa-miR-199a-5p (red) and hsa-miR-199a-3p (blue) targets and all genes (grey) in the i199 (**B**) and i199-KTN1 (**E**) cells, in function of $\log_2$ GFP expression in the same cells; see Methods about PAGODA normalization and calculation details. **D.** Total expression ($\log_2$ sum of TPMs) of 1000 random genes in the i199 (lower line) and i199-KTN1 (upper line) cells. **C,F.** Mean Pearson pairwise correlation coefficients for miRNA targets in function of GFP expression in i199 (**C**) and i199-KTN1 (**F**) cells. Mean from 50 calculation evaluations of random selection of 100 genes is shown as grey line. Means were calculated from the two hundred cells with GFP expression closest to a specific expression level (**C,D,F**). For **A** and **D** panels standard deviations are shown, for **B,C,E** and **F** standard error. Similar to Figure 3.8, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

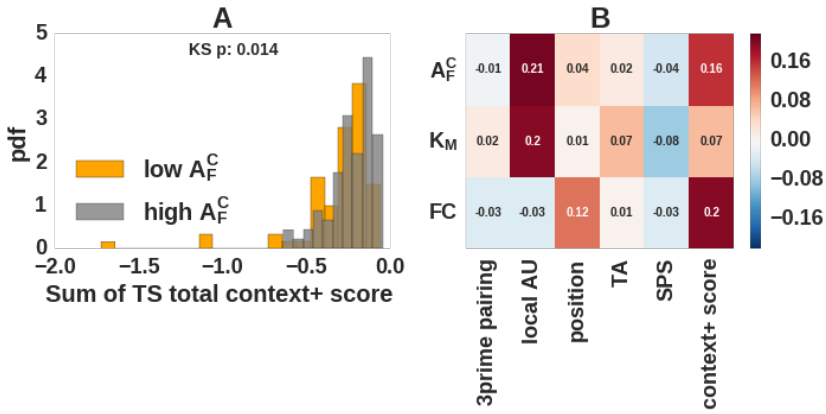


Figure 3.17: **Appendix Figure S12. hsa-miR-199a-3/5p targets A_F^C correlate with binding site properties. A.** Low A_F^C targets have lower Target-Scan scores compared to high A_F^C targets. For each gene, we calculated the mean of the A_F^C values inferred from the i199 and i199-KTN1 data. We also calculated total Target-Scan scores for each gene by summing the prediction scores for the two miRNAs. We took the union of the 40 targets with with lowest A_F^C for the two miRNAs (72 targets, as some were targeted by both miRNAs) and the similar list of targets with the highest A_F^C and compared their Target-Scan scores. **B.** Spearman correlation of A_F^C, K_M and FC to site properties. Binding site properties used to calculate Target-Scan context+score for all targets of the two miRNAs which have only one binding site (of either of the two miRNAs, about 70% of targets) are used in this correlation. Note that features that form Target-Scan and the context+score have minus values, as opposite to MIRZA-G-C score and its partial elements. Similar to Figure 3.10, but using TargetScan 6.2 instead of MIRZA-G-C-predicted targets.

3.5.2 Appendix tables

Nr	GO name	GOMFID	$P-$value	ExpCount	Count	Size
1	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	GO:0016810	0.00376701	1.0070922	4	4
2	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	GO:0001077	0.003860534	2.0141844	6	8
3	RNA polymerase II transcription factor activity, sequence-specific DNA binding	GO:0000981	0.03697759	4.2801418	8	17
4	syntaxin binding	GO:0019905	0.050522247	1.0070922	3	4
5	histone deacetylase activity	GO:0004407	0.062719265	0.5035461	2	2
6	monooxygenase activity	GO:0004497	0.062719265	0.5035461	2	2
7	deacetylase activity	GO:0019213	0.062719265	0.5035461	2	2
8	heme binding	GO:0020037	0.062719265	0.5035461	2	2
9	tau protein binding	GO:0048156	0.062719265	0.5035461	2	2
10	sequence-specific DNA binding	GO:0043565	0.065069175	6.2943262	10	25
11	helicase activity	GO:0004386	0.069707031	1.7624113	4	7
12	transcription regulatory region sequence-specific DNA binding	GO:0000976	0.097208599	5.035461	8	20

Table 3.1: **Appendix Table S1. "Molecular function" GO categories enriched in targets with low A_F^C.** The union of the 40 targets with the lowest A_F^C for each of the two miRNAs (mean over i199 and i199-KTN1 data, 74 targets in total) was used as foreground set. As a background the joint list of targets (all for which A_F^C values were calculated) for 3p and 5p arm (present in both i199 and i199-KTN1 data sets; 301 in total) was used. 15 of the 74 targets are annotated with GO terms 2,3,5,10-12, related to transcription regulation and DNA binding.

Nr	GO name	GOMFID	Pvalue	ExpCount	Count	Size
1	organic anion transmembrane transporter activity	GO:0008514	0.01355234	0.7234043	3	3
2	transmembrane transporter activity	GO:0022857	0.02635304	2.6524823	6	11
3	transferase activity	GO:0016772	0.03671596	5.787234	10	24
4	substrate-specific transporter activity	GO:0022892	0.04285169	3.6170213	7	15
5	carbohydrate derivative binding	GO:0097367	0.05152047	11.0921986	16	46
6	NAD+ kinase activity	GO:0003951	0.05749476	0.4822695	2	2
7	diacylglycerol kinase activity	GO:0004143	0.05749476	0.4822695	2	2
8	organic acid transmembrane transporter activity	GO:0005342	0.05749476	0.4822695	2	2
9	amino acid transmembrane transporter activity	GO:0015171	0.05749476	0.4822695	2	2
10	antiporter activity	GO:0015297	0.05749476	0.4822695	2	2
11	adrenergic receptor binding	GO:0031690	0.05749476	0.4822695	2	2
12	metal ion transmembrane transporter activity	GO:0046873	0.05749476	0.4822695	2	2
13	catalytic activity	GO:0140103	0.05749476	0.4822695	2	2
14	adenyl nucleotide binding	GO:0030554	0.07966644	9.8865248	14	41
15	anion binding	GO:0043168	0.08369943	13.5035461	18	56
16	purine ribonucleotide binding	GO:0032555	0.085241	10.8510638	15	45
17	ion transmembrane transporter activity	GO:0015075	0.08612716	1.1732852	3	5
18	ATP binding	GO:0005524	0.08957296	9.1631206	13	38
19	calmodulin binding	GO:0005516	0.09252429	1.2056738	3	5
20	cation transmembrane transporter activity	GO:0008324	0.09252429	1.2056738	3	5
21	tubulin binding	GO:0015631	0.09660753	2.6524823	5	11

Table 3.2: **Appendix Table S2. "Molecular function" GO category analysis for high A_F^C targets.** The union of the 40 targets with the highest A_F^C for each of the two miRNAs (mean over i199 and i199-KTN1 data, 77 targets in total) was used as foreground set. As a background the joint list of targets (all for which A_F^C values were calculated) for 3p and 5p arm (present in both 199 and i199-KTN1 data sets; 301 in total) was used.

Chapter 4

Bayesian inference of the gene expression states from single-cell RNA-seq data

Jérémie Breda[1,2], Mihaela Zavolan[1,2], Erik van Nimwegen[1,2]

1. Biozentrum, University of Basel, Basel, Switzerland
2. Swiss Institute of Bioinformatics, Basel, Switzerland

Nature Biotechnology 39 (2021) 1008–1016

Abstract

Despite substantial progress in single-cell RNA-seq data analysis methods, there is still little agreement on how to best to normalize such data. Starting from basic requirements such as that inferred expression states should correct for both biological and measurement sampling noise, and that changes in expression should be measured in terms of fold-changes, we here derive a Bayesian normalization procedure called Sanity (SAmpling Noise corrected Inference of Transcription activitY) from first principles. Sanity estimates expression values and associated errors bars directly from raw unique molecular identifier counts without any tunable parameters. Using simulated and real scRNA-seq datasets, we show that Sanity outperforms other normalization methods on downstream tasks such as finding nearest-neighbors cells, and clustering cells into subtypes. Moreover, we show that by systematically overestimating the expression variability of low expressed genes and by introducing spurious correlations through mapping the data to a lower-dimensional representation, other methods yield severely distorted pictures of the data.

4.1 Introduction

In the past decade, much effort has been invested in adapting methods for quantifying transcriptome and epigenome states on a genome-wide scale to the single-cell level. This has led to a large number of new methods that are starting to make it possible to track the states of single cells across tissues and embryos as they are developing [Picelli et al., 2013, Hashimshony et al., 2012, Macosko et al., 2015, Klein et al., 2015, Buenrostro et al., 2015, Cusanovich et al., 2015, Rotem et al., 2015, Smallwood et al., 2014, Nagano et al., 2013, McKenna et al., 2016, Kalhor et al., 2018, Frieda et al., 2017, Frei et al., 2016, Raj et al., 2018, Spanjaard et al., 2018, Wagner et al., 2018, Angermueller et al., 2016, Clark et al., 2018, Adamson et al., 2016, Dixit et al., 2016, Jaitin et al., 2016, Datlinger et al., 2017]. It is widely expected that these methods will revolutionize our understanding of the ways in which cell fate and cell identity are regulated, and large consortia are being formed with the aim to comprehensively chart single-cell landscapes in model organisms [Regev et al., 2017, Rajewsky et al., 2020].

To fulfill the promise of these single-cell measurement technologies, it will be crucial that computational methods are available to unambiguously determine what the raw measurements say about the states of individual cells. We not only want to be able to integrate results of single-cell RNA-seq (scRNA-seq) measurements from different labs and protocols, but also with results of different measurement technologies such as FISH (e.g. [Raj et al., 2008]). To make this possible, the expression values that we extract from scRNA-seq data should correspond to physically meaningful quantities that can be directly compared with measurements of

the same quantities made with other experimental methods. In addition, the estimated values of these concrete physical quantities should follow directly from the experimental data with as small a number of additional assumptions as possible, and not depend on arbitrary parameters that the user can set at will. Moreover, to be able to determine when different measurements are mutually consistent, estimates should be accompanied by error bars.

However, although there has been a veritable explosion of scRNA-seq analysis tools in recent years, little attention has been given to satisfying these objectives. Instead of a few methods that estimate quantities with clear physical interpretation in a transparent manner, scientists are faced with a large number of *ad hoc* methods that apply complex transformations to the data to perform combinations of tasks including imputation/normalization, clustering, dimensionality reduction, pseudotime and trajectory inference, and visualization. These methods often have many tunable parameters, produce outputs in abstract spaces that lack clear biological meaning, and are often even stochastic, giving varying outputs even when run on the same data with the same parameters. For example, the popular t-SNE [Van Der Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018] visualization tools, are both stochastic, highly-dependent on parameter settings, and position cells in a space whose dimensions lack biological interpretation.

Here we focus on the basic task of normalization/imputation of single-cell gene expression states from raw scRNA-seq transcript counts. Using only minimal assumptions we derive from first principles a Bayesian method that corrects not only for the finite sampling associated with the capture and sequencing of mRNAs, but also for the Poisson noise inherent in the gene expression process itself. Our method, which we call Sanity (SAmpling Noise corrected Inference of Transcription activitY) is deterministic, has no tunable parameters, and provides error-bars for all its estimates.

We compare Sanity with a selection of popular methods for imputation/normalization from the recent literature ([Eraslan et al., 2019, Lun et al., 2016, van Dijk et al., 2018, Huang et al., 2018, Li and Li, 2018, Hafemeister and Satija, 2019, Lopez et al., 2018], see methods) and show that only Sanity can effectively remove Poisson sampling fluctuations to infer the true variation in gene expression of each gene across cells. In addition, we show that all other methods we tested introduce severe distortions of the data such as inducing strong correlations between expression estimates and total Unique Molecule Identifier (UMI) count of cells, or inferring strong co-expression between large numbers of genes when none is evident in the data. Finally, we show that the expression levels estimated by Sanity outcompete those of other methods in downstream analysis tasks such as finding nearest-neighbor cells and clustering.

4.2 Results

Sanity's approach, which is detailed in the online methods, is summarized in Fig. 4.1. Although it is tempting to consider the gene expression state of a cell to simply correspond to the vector of its mRNA counts, these mRNA counts will exhibit Poisson fluctuations from cell to cell, even if the rates of transcription and mRNA decay were constant across cells and time. We thus argue that changes in expression state should only reflect changes in transcription and decay rates of mRNAs, and correct for the intrinsic noise in gene expression. The crucial insight is that, even if transcription and mRNA decay rates vary with time in an arbitrary way in a given cell, the mRNA count m_g of each gene g is still just a Poisson sample of a single effective 'transcription activity' a_g, which is a weighted average of its recent transcription and mRNA decay rate in the cell (Fig. 4.1a and b). Sanity represents the expression state of a cell by a vector of transcription quotients α_g corresponding to these relative transcription activities (Fig. 4.1b and c). As shown in the online and supplementary methods, the probability of the raw UMI counts of a cell given its transcription quotients is a product of Poisson distributions (Fig. 4.1d).

In order to infer the log transcription quotients (LTQ) of each gene in each cell from the UMI counts, Sanity makes as few prior assumptions about how LTQs might vary across genes and cells. In particular, it only assumes that, for each gene g, the distribution of its LTQs across cells can be characterized by an unknown mean μ_g and variance v_g. Given this, the entire inference procedure follows from first principles, without any tunable parameters.

As detailed in the online methods, we use 7 real and 2 simulated scRNA-seq datasets to compare Sanity's performance with those of two basic normalization methods that simply log-transform raw or normalized UMI counts (called Raw-Counts, and TPM, respectively), and 7 other recently proposed normalization methods (DCA [Eraslan et al., 2019], Deconvolution [Lun et al., 2016], MAGIC [van Dijk et al., 2018], SAVER [Huang et al., 2018], scImpute [Li and Li, 2018], sctransform [Hafemeister and Satija, 2019], and scVI [Lopez et al., 2018]).

4.2.1 Sanity accurately corrects for Poisson fluctuations to identify true variance in gene expression

A key aim of Sanity's normalization is to correct for both biological and technical sampling noise to quantify the true biological variation in gene expression across cells. Testing this is challenging because the true expression variability of each gene is generally unknown. To address this issue we first analyzed a carefully designed study of mouse embryonic stem cells (ESCs) from *Grün et al* [Grün et al., 2014] in which not only scRNA-seq measurements were taken for cells cultured in both 2i and serum conditions, but the same measurement protocol was also applied

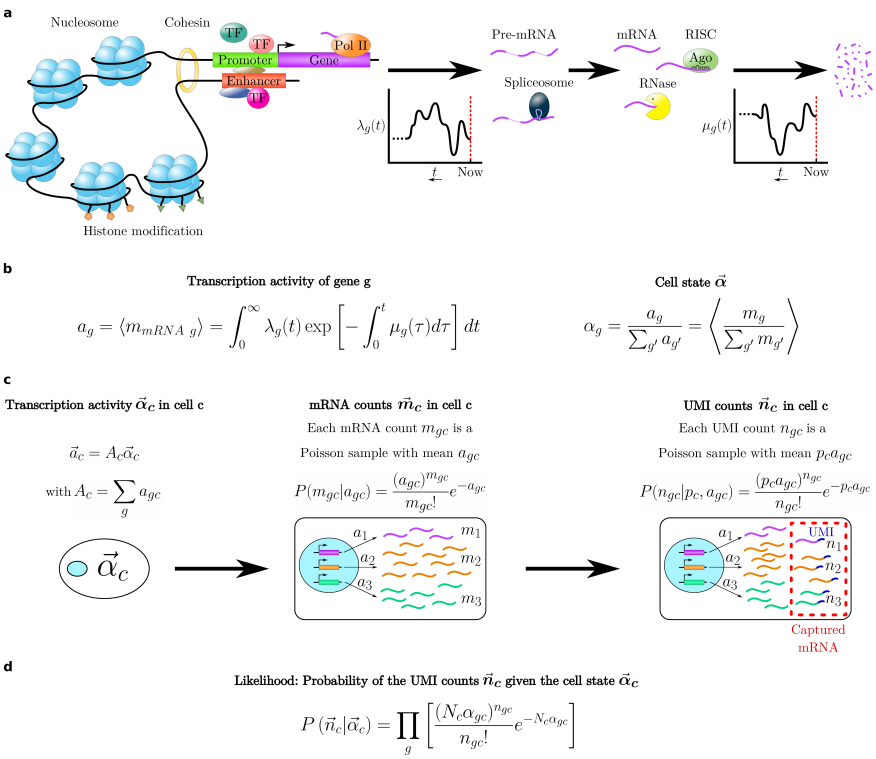


Figure 4.1: Summary of the Sanity approach. **a**: Cartoon of the flow of causality from the physical state of the cell to gene expression patterns. The concentrations of transcription factors (TFs), chromatin modifiers, and other regulatory factors determine changes in chromatin state, 3D organization of the chromosomes, binding and unbinding of TFs to promoters and enhancers, and so on. These determine the time-dependent rate $\lambda_g(t)$ at which gene g was described a time t in the past. Similarly, the concentrations of microRNAs, RNAses and other RNA binding proteins determine the time dependent rate $\mu_g(t)$ at which mRNAs of gene g decayed a time t in the past. **b**: The 'transcription activity' a_g of gene g is defined as the expected number of mRNAs and is a weighted average of its transcription and decay rates in the past. We define the expression state of the cell as the vector $\vec{\alpha}$ of relative transcription activities of all genes. **c**: Logical flow from expression state $\vec{\alpha}_c$ to observed UMI counts $\vec{n}_c$. The expression state $\vec{\alpha}_c$ and total transcription activity A_c determine the transcription activities a_{gc}. For each gene g, the probability $P(m_{gc}|a_{gc})$ to have m_{gc} mRNAs is a Poisson distribution with mean a_{gc}. Assuming each mRNA in cell c has a probability p_c to be captured and sequenced, the probability $P(n_{gc}|p_c, a_{gc})$ to obtain n_{gc} UMIs is a Poisson distribution with mean $p_c a_{gc}$. **d**: The probability to obtain the UMI counts $\vec{n}_c$ given the cell state $\vec{\alpha}_c$ is a product over genes of Poisson distributions with means $N_c \alpha_{gc}$, where N_c is the total UMI count in cell c.

to single-cell equivalent *aliquots* from pooled RNA. The expression variation in these aliquots thus solely derives from technical sampling noise. In addition, the ESCs are highly homogeneous so that little true expression variation is expected for ESCs in the same condition.

Figure 4.2a shows box-whisker plots of the distributions of coefficients of variation (CV) across genes for each of the 4 datasets, as calculated from the expression estimates of each of the normalization methods (except for sctransform, which does not report estimated expression values). Analogous results using standard-deviation in LTQ (which is equivalent to CV when CV is small, see Suppl Text 1) are shown in Suppl. Fig.4.6.

Ideally the methods should infer that there is no true variability at all for the aliquots, and relatively little variability for the ESCs. However, although all methods infer that CVs are slightly larger in serum than 2i, which is in line with previous analysis [Grün et al., 2014], most methods infer substantial variability for most genes. In particular, methods that do not correct for Poisson noise (RawCounts, TPM, Deconvolution, and scImpute) infer CVs of larger than 0.5 for the large majority of genes in both cells and aliquots. In contrast, the CVs that Sanity infers are at least twofold lower than those of all other methods, and only Sanity correctly infers that there isn no expression variability in the aliquots, i.e. with CVs less than 10% for almost all genes.

There is no reason to expect that CVs in expression should correlate with mean expression, and in bulk RNA-seq there is indeed no correlation between mean log-expression and variance in log-expression across conditions (Suppl. Fig. 4.7). However, at the single-cell level, the intrinsic Poisson fluctuations will add a term $1/\sqrt{\text{mean}}$ to the CV, as is well appreciated in the scRNA-seq literature, e.g. [Brennecke et al., 2013]. Thus, systematic correlations between CV and mean of normalized expression levels reflect to what extent a method has failed to correct for Poisson sampling noise. Figure 4.2b shows scatter plots of CV against mean expression for all methods and we see that, with the exception of Sanity and MAGIC, all other methods show a strong negative correlation between CV and mean, indicating that Poisson sampling noise dominates the observed variability for all but the highest expressed genes. These observations apply to all datasets (Suppl. Fig. 4.8), including the simulated dataset that we discuss next.

To more directly test the accuracy with which different methods estimate the expression variance of each gene, we constructed a simulated dataset for which the true mean and variation in LTQ across cells is known for each gene (see Online Methods and Suppl. Fig. 4.9). Comparing the true CVs of each gene with those inferred by each method (Suppl. Fig. 4.10) shows that only Sanity, and to a lesser extent Saver, exhibit a good correlation between true and inferred CVs. A comparison of true and inferred variances in LTQs confirms this overall picture (Suppl. Fig. 4.11). Notably, for all methods except for Sanity, the Poisson noise causes the inferred CVs of low expressed genes to be systematically higher than

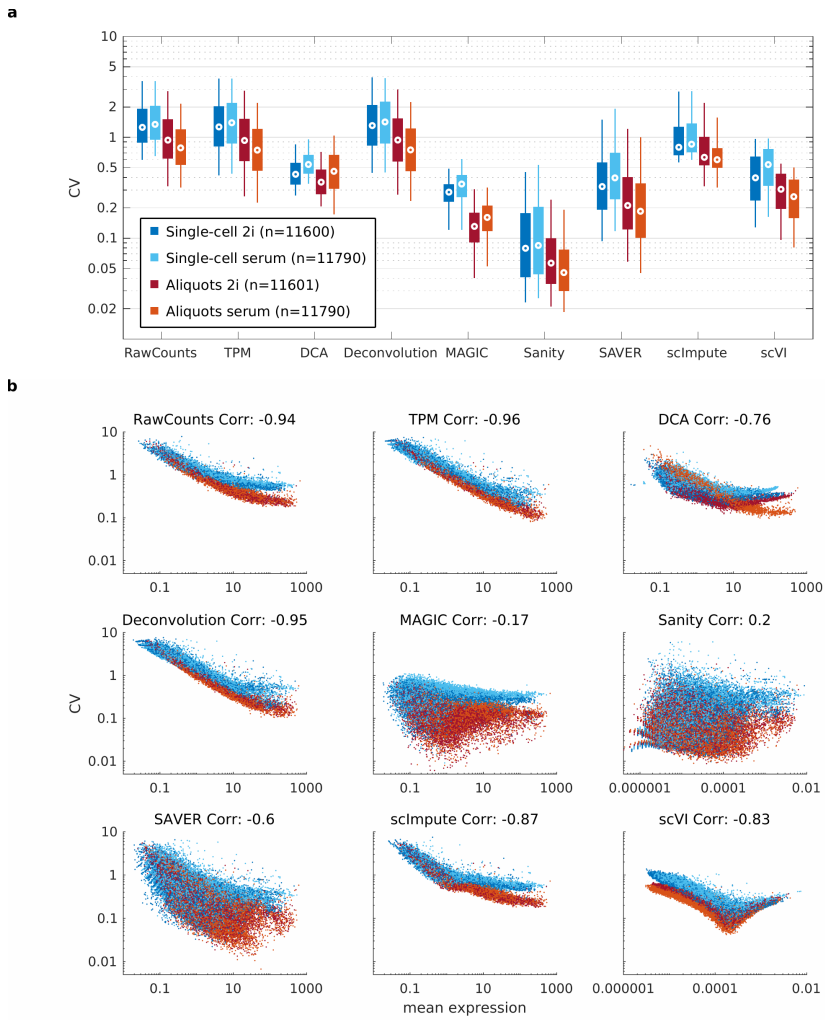


Figure 4.2: Effects of Poisson fluctuations on gene expression variance. **a**: Box-whisker plots showing the median (circle) as well as the 5th, 25th, 75th, and 95th quantiles of the distribution of coefficients of variation (CV) of gene expression levels across genes, for each of the 4 datasets (see legend) as inferred by each of the normalization methods. **b** Scatter plots of CV (standard-deviation divided by mean) against mean expression for all genes in each of the 4 datasets (colors as in panel a) as inferred by each of the normalization methods. The Pearson correlation coefficient between log CV and log mean is shown on top of each plot. The axes are shown on logarithmic scales and are kept similar across panels, except for Sanity and scVI where the mean expression values are on a very different scale from those of the other methods.

their true CV, resulting in an almost complete loss of correlation between true and inferred CV across genes for most methods. For very low expressed genes the

expression data is so sparse that it is only possible to estimate an upper bound on expression variability (see Suppl. Text 1) and Sanity conservatively infers that the true expression variability is low, so that these genes will not significantly contribute to most downstream analyses.

In summary, Sanity is the only normalization method that can reliably correct for the Poisson sampling noise to estimate the true expression variability of each gene.

4.2.2 The accuracy of gene expression estimates strongly depends on the depth of coverage

The gene expression measurement noise is expected to scale inversely proportional to absolute expression, i.e. for a gene with $\langle n \rangle$ expected UMI in a cell the Poisson noise will cause the measured log-expression $\log(n)$ of a gene to differ from the true log-expression $\log(\langle n \rangle)$ by a term of order $1/\sqrt{n}$. We thus used the same simulated dataset to compare the accuracy of the gene expression estimates of the different methods methods as a function of depth of coverage. In particular, we stratified all genes into bins according to their absolute expression (average number of UMI per cell) and calculated the accuracy of various expression estimates for each method and each bin (Fig. 4.3).

While most methods accurately estimate mean log-expression levels for genes with at least 0.1 UMIs per cell, DCA, scVI, sctransform, and scImpute never do (Fig. 4.3a). Second, although Sanity is essentially the only method that can accurately estimate the true variance in log-expression levels across cells, even Sanity can only reliably estimate the true variance in LTQ for genes that have at least 1 UMI per cell on average (Fig. 4.3b). Third, the Pearson correlations between true and estimated log fold-changes quantify how accurately each method identifies in which cells a gene is highest and lowest expressed (Fig. 4.3c). We observed that Pearson correlations systematically increase with absolute expression, with Sanity performing best at each expression level, followed closely by TPM, Deconvolution and SAVER. In contrast, the log fold-changes predicted by MAGIC, DCA and scVI show almost no correlations with the true log fold-changes, even for highly expressed genes, suggesting that these method systematically distort expression levels. However, even for the best methods, correlations are only consistently high for genes with at least 1 UMI per cell, and consistently low for genes with less than 0.1 UMI per cell.

As discussed in Suppl. Text 1, with current capture efficiencies the vast majority of genes have less than 1 UMI per cell (Suppl. Fig. 4.31). As accurate estimates of expression levels are only guaranteed for genes with at least 1 UMI per cell (Fig. 4.3), this implies accurate estimates of expression patterns for only a few hundred genes. Consequently, if it were possible to substantially raise capture and sequencing efficiencies, the number of genes for which we would be able to

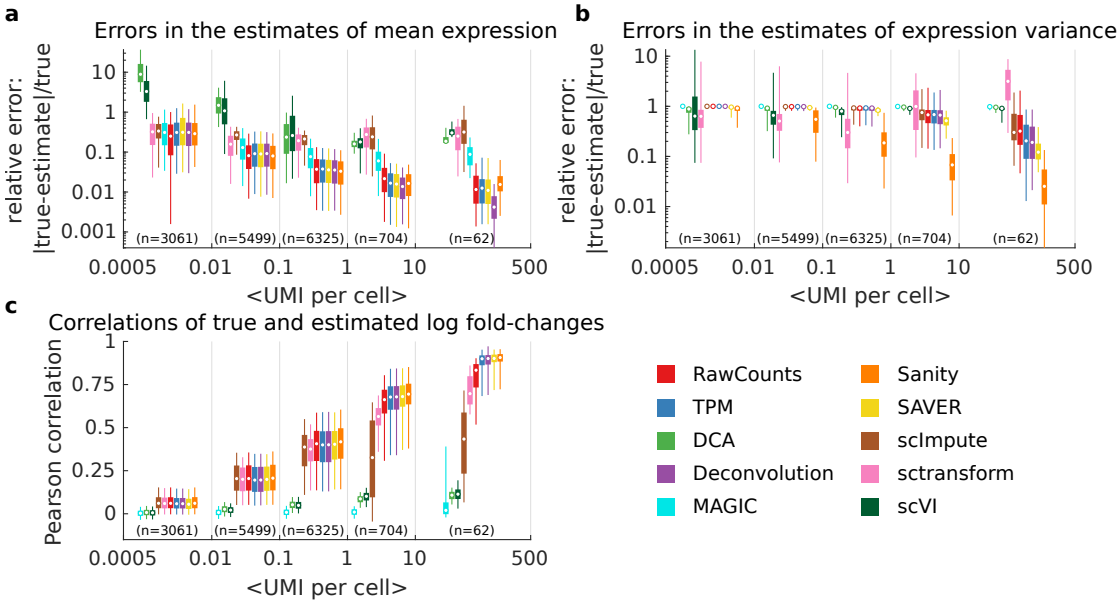


Figure 4.3: Accuracy of the gene expression estimates as a function of depth of coverage. Genes were stratified into 5 bins of absolute expression (in average number of UMI per cell) and for each bin the distribution of relative errors in estimated mean log-expression (panel **a**), estimated variance in log-expression (panel **b**), and Pearson correlations between true and estimated log fold-changes across cells (panel **c**) were calculated for each method (different colors, see legend). The distributions are shown as box-whisker plots showing the median, inter-quartile range, 5 percentile, and 95 percentile for the genes in each expression bin. Note that the vertical axes are shown on a logarithmic scale for the top two panels. Methods are sorted from right to left in approximate order of their accuracy in each panel.

obtain accurate expression estimates could be dramatically increased (Suppl. Fig. 4.31).

4.2.3 Many normalization methods introduce spurious correlations

Due to variation in cell size, mRNA capture efficiency, and sequencing depth, the total number of UMIs can fluctuate significantly from cell to cell. Therefore, most scRNA-seq processing methods normalize expression levels for the total number of mRNAs (i.e. UMIs) that were captured from a given cell. The simple TPM procedure does so by dividing the observed counts for each gene by the total UMI count of the cell, and Deconvolution accomplishes the same normalization using a more sophisticated approach. With the exception of RawCounts and scImpute, all other methods normalize for total UMI count.

If the normalization for total UMI count were successful, we would expect no

systematic correlation between inferred expression levels and total UMI counts across cells for most genes. However, this is not what we observe. For each method and gene, we calculated the Pearson correlation between the inferred log-expression levels and log total UMI counts. Using the Zeisel dataset as an example, Fig. 4.4a-b shows the distribution of Pearson correlations, as well as raw scatters of the normalized expression levels as a function of log total UMI count for one example gene (*Zbed3*).

As expected, because RawCounts and scImpute do not normalize for total UMI count N_c, for these methods most genes show a positive correlation between the inferred expression levels and $\log(N_c)$. In contrast, the simple TPM method, Deconvolution, and especially Sanity and sctransform, successfully remove this correlation. However, although DCA, SAVER, MAGIC and scVI also intend to normalize for total UMI counts, their normalized expression levels show even stronger correlations with $\log(N_c)$ than the non-normalized RawCounts. The scatters with the inferred expression levels for the gene *Zbed3* as a function of $\log(N_c)$ illustrate how dramatically some normalization methods transform the input data. The RawCounts show that this gene has fairly low expression, with 0 or 1 UMIs observed in most cells, and with a slightly higher chance to observe 1 or 2 UMIs when the total UMI count N_c is larger. However, DCA, MAGIC, SAVER, and scVI completely transform this input data into a scatter of continuously varying expression levels that either correlate negatively (DCA, SAVER, scVI) or strongly positively (MAGIC) with total UMI count. These observations again generalize to all other datasets as shown in Suppl. Fig. 4.12.

In many studies, systematic analysis of the co-expression of pairs of genes is used to identify co-regulated pathways or regulatory modules. For such applications, it is thus crucial that the pairwise correlations of the expression profiles accurately reflect the co-expression evidence in the data. To investigate this, we calculated Pearson correlations of the normalized log-expression levels of all pairs of genes, and then compared these pairwise correlation coefficients across the various methods, using the Baron dataset as an example (Fig. 4.4c-g). The pairwise correlations by-and-large agree between Sanity and the simple TPM method (Fig. 4.4c), and this agreement is also observed for Deconvolution and sctransform (Suppl. Fig. 4.13). Although Sanity and scImpute also by and large agree on which pairs of genes are most strongly positively or negatively correlated (Fig. 4.4d), scImpute predicts moderately positive correlations for many gene pairs for which Sanity predicts no correlation at all. This behavior results from scImpute not normalizing for total UMI count, and is indeed also observed for RawCounts (Suppl. Fig. 4.13).

A very different pattern is observed for the comparison of Sanity with MAGIC (Fig. 4.4e). For many of the pairs of genes for which Sanity infers no co-expression, i.e. zero correlation, MAGIC infers a broad range of correlations running from almost perfect anti-correlation, to perfect correlation. To further investigate this,

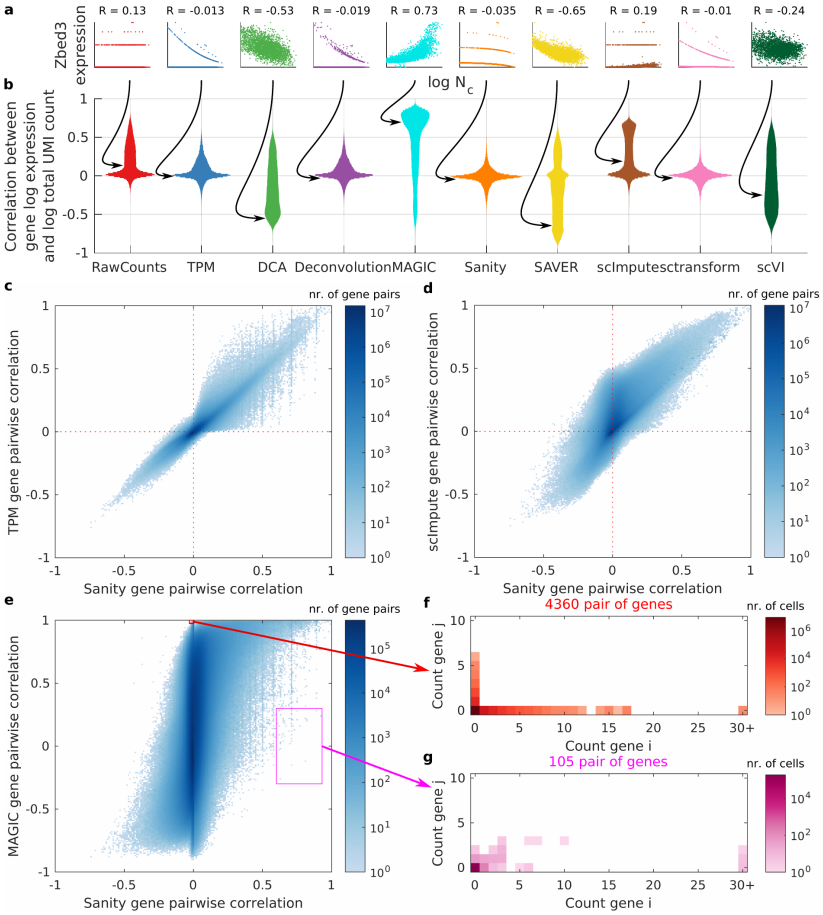


Figure 4.4: Correlations between inferred gene expression levels and library size, and between pairs of genes. **a**: Scatter plots of the normalized log-expression $\log(e_c)$ of the example gene (*Zbed3*) versus the logarithm of the total UMI count $\log(N_c)$ across cells for each method. The Pearson correlation of the dependence is shown above each panel. **b**: Violin plots of the distribution of correlation coefficients between $\log(e_c)$ and $\log(N_c)$ for all genes, for the *Zeisel* dataset. Each color corresponds to a method, indicated below each plot. **c-e**: Density plots of Pearson correlations for all pairs of genes as inferred by *Sanity* (x-axis) against the correlations inferred by *TPM*, *scImpute*, and *MAGIC* (y-axis). The color scale shows the density in $\log_{10}$ number of gene pairs and zero counts are shown in white. The red and the magenta rectangles in panel e indicate the pairs of genes with correlation above 0.975 for *MAGIC* and between -0.03 and 0.005 for *Sanity* (red), and all pairs with correlation between -0.3 and 0.3 for *MAGIC* and between 0.6 and 0.93 for *Sanity* (magenta). **f**: 2-dimensional histogram of counts per cell summed over the 4360 pairs of genes from the red rectangle in panel e. The height of the histogram is shown in $\log_{10}$ as a color and zero counts are shown in white. **g**: Analogous 2-dimensional histogram of counts for the 105 pairs of genes from the magenta rectangle in panel e.

we focused on a subset of 4360 pairs of genes within the red rectangle of Fig. 4.4e, for which MAGIC predicted nearly perfect correlation and Sanity almost none. Summing across all 4360 pairs of genes and all cells, we found there was not a single example for which both genes in a pair were observed in the same cell (Fig. 4.4f). That is, although MAGIC infers that these 4360 pairs of genes are almost perfectly co-expressed, *none* of them are ever observed to be present at the same time in *any* cell. In contrast, for the small set of pairs for which Sanity infers co-expression whereas MAGIC does not, we do generally find evidence of co-expression (Fig. 4.4g). This same pattern is observed for the comparisons of Sanity's pairwise correlations with those of DCA, SAVER, and scVI (Suppl. Fig. 4.14). That is, these methods all infer large numbers of highly correlated or anti-correlated pairs of genes, whereas there is no evidence at all of co-expression in the raw counts of these pairs. Consistent with these observations, these methods show very wide distributions of pairwise correlations on each dataset, whereas correlations are highly peaked around zero for Sanity, TPM, Deconvolution, and sctransform (Suppl. Fig. 4.15). Moreover, although our simulated dataset contains no correlations by construction, DCA, MAGIC, scVI, and to a lesser extent SAVER also predict a wide range of correlations on this data (Suppl. Fig. 4.16).

We believe that these pervasive spurious correlations result from the fact that these methods map the expression data to a lower dimensional manifold. Indeed, if we project the TPM-normalized results from the simulated data on the first n PCA components, the amount of spurious correlations systematically increases with decreasing n (Suppl. Fig. 4.17). Comparison of Suppl. Figs. 4.16 and 4.17 shows that the amount of spurious correlation in SAVER's results is equivalent to projecting on the first $100 - 200$ PCs, the first $20 - 30$ PCs for DCA and scVI, and the first $5 - 10$ PCs for MAGIC.

4.2.4 Sanity outperforms other methods on identifying nearest-neighbor cells

Many downstream scRNA-seq analyses including clustering and trajectory reconstruction, require estimating the distances between cells in gene expression space. In particular, many methods involve identifying the k nearest-neighbors of each cell with the most similar expression profiles (with k typically in the range of $3 - 30$). Assessing the accuracy of different methods in identifying nearest-neighbor cells on real data is challenging because it is not known which cells are truly nearest-neighbors. We thus created a simulated dataset in which cells are distributed along a tree that was constructed by performing a branched random walk through gene expression space, i.e. setting the true LTQs of each cell equal to those of the previous cell plus a small random perturbation to the LTQ of each gene (see Suppl. Methods).

For each method, we calculated the Euclidean distances between the normalized

log-expression vectors of all pairs of cells, and determined the k nearest-neighbors of each cell. For Sanity we also estimated cell-to-cell distances using a Bayesian method that incorporates Sanity's error bars, which automatically causes genes with large error bars ϵ_{gc} to contribute less to the distance estimate (Suppl. Methods). For each method we then calculated the fraction of predicted k nearest-neighbors that belong to the set of true k nearest-neighbors, as a function of k (Fig. 4.5a).

Sanity clearly performs best in identifying the k nearest-neighbors, but when its error bars are ignored the performance is much reduced, highlighting the value of incorporating error-bars. This reduction in performance is due to the noisy estimates of the LTQs of low expressed genes because, if we calculate distances based only on the genes with at least 1 UMI per cell on average, Sanity's performance without error bars is dramatically improved, approaching the performance incorporating error bars for large k (Fig. 4.5b). Other normalization methods, e.g. TPM and sctransform, also perform much better when distances are only estimated from genes with at least 1 UMI per cell. In contrast, the performance of scVI and DCA is not sensitive to excluding low expressed genes, suggesting that for these methods the expression levels of low expressed genes are effectively determined by the expression levels of high expressed genes. Notably, whereas DCA and scVI performed poorly on previous tests concerned with the accuracy of inferred gene expression levels, here they are the best performing methods after Sanity, and also perform well at estimating distances between all pairs of cells (Suppl. Figs. 4.18 and 4.19). This shows that these methods are optimized to correctly estimate distances between cells at the expense of severely distorting the expression patterns of individual genes.

To give a visual impression of the accuracy with which different methods are able to capture the local structure in the data, Suppl. Fig. 4.20 shows t-SNE visualizations of the matrices of true cell-to-cell distances and cell-to-cell distances as estimated by each of the methods. It is notable that, even though the data corresponds to a complex tree structure of 149 branches with 13 cells each, Sanity's estimates of the cell-to-cell distances allow a reasonably accurate reconstruction of this complex structure.

4.2.5 Sanity outperforms other methods on clustering cells into subtypes

One of the main applications of scRNA-seq is to identify (novel) cell types and this is generally done by clustering cells based on their gene expression patterns. For six of our test datasets, the corresponding study reported an annotation of cell types, that was typically obtained by combining automated clustering with analysis of marker gene expression and hand curation using prior knowledge. Taking the Zeisel dataset as an example [Zeisel et al., 2015], Fig. 4.5c visualizes the clus-

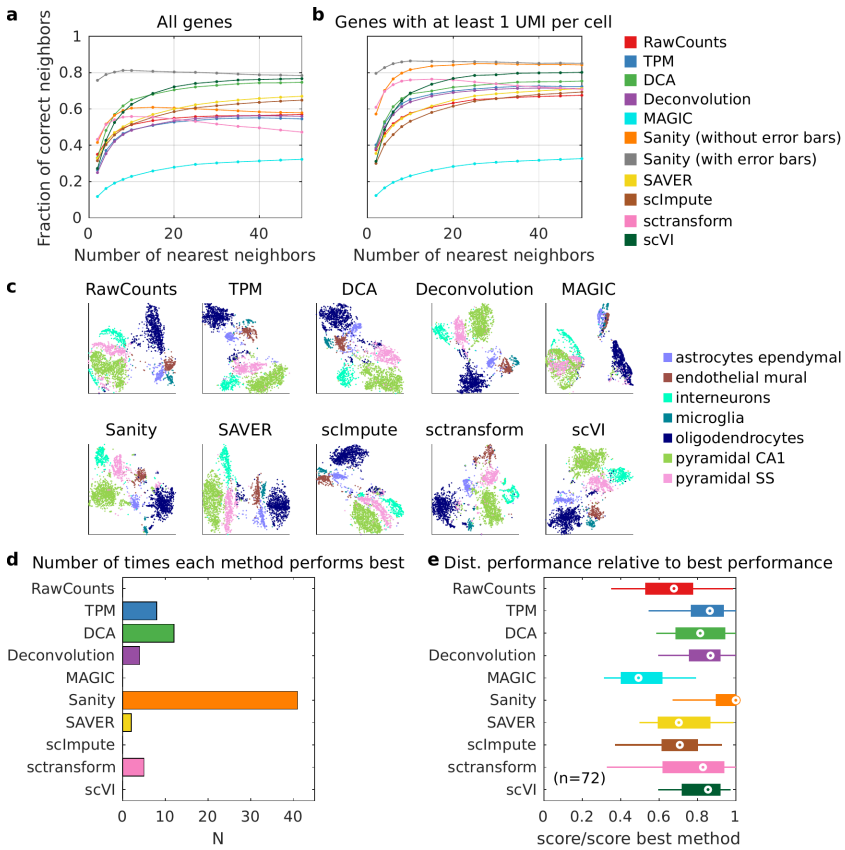


Figure 4.5: Accuracy of the k nearest-neighbor and clustering predictions. **a**: For each method we calculated the Euclidean distances between the log-expression profiles of all pairs of cells to predict the nearest-neighbors of each cell, on the simulated dataset for which cells lie along a branched random walk in gene expression space. The curves show the fraction of predicted k nearest-neighbors for each method (colors, see legend) that are members of the set of true k nearest-neighbors, as a function of k. **b**: As in panel A but now calculating distances using only highly expressed genes (at least 1 UMI per cell on average). **c**: Each panel shows a t-SNE visualization of the Zeisel dataset using the normalized gene expression values of the method indicated at the top of the panel. Each point represents a cell and is colored according to the cell type annotation given in [Zeisel et al., 2015]. **d**: Similarity scores between the annotated and predicted clustering were calculated, for each method, across 72 combinations of 6 annotated datasets, 3 clustering algorithms, and 4 similarity metrics. The bars show, for each method, the number of combinations on which it performed best. **e**: For each method m, the distribution (across the 72 combinations) of the ratio s_m/s_* of its similarity score s_m relative to the similarity score s_* of the best performing method (on that combination) is shown as a box-whisker plot indicating 5the percentile, first quartile, median, third quartiles, and 95th percentile.

tering structure implied by the different methods by applying the popular t-SNE algorithm [Van Der Maaten and Hinton, 2008] to the normalized expression values of each method. Although it is well-known that, beyond reasonably conserving which cells are nearest-neighbors, it is difficult to interpret these visualizations, the visualization does suggest that there is considerable disparity between normalization methods. In particular, Sanity, TPM, and Deconvolution appear to separate the cell types more reliably than MAGIC, RawCounts, and scImpute, and similar observations can be made on the other datasets (Suppl. Figs. 4.21-4.25).

Rigorously benchmarking the performance of the normalization methods on clustering is challenging because the ground truth is again not known. While the provided reference annotations are likely reasonable, it is by no means clear that these annotations are optimal. In addition, the clustering performance will also depend on what clustering algorithm is used, and even what similarity measure is used to compare clusterings. We thus chose to assess the quality of each normalization method by its performance across all 6 datasets using 3 different clustering algorithms (K-means [Lloyd, 1982], Ward [Ward, 1963], and Louvain [Blondel et al., 2008]) and using 4 different similarity measures (Suppl. Methods), giving 72 comparisons of similarity scores across methods (Suppl. Fig. 4.26). To summarize these results, we calculated the number of times each method was the best performing method (Fig. 4.5d). In addition, we a calculated how close each method comes to the best performing method across the 72 combinations (Fig. 4.5e). Sanity clearly outperforms the other methods, being the best performing method on more than half of the combinations, and scoring close to the best performing method on the large majority of combinations. TPM, Deconvolution, DCA, and scVI also perform robustly, typically scoring within 10% of the best method.

As a final example of downstream analysis we tested the ability of the normalized expression values to identify genes that are significantly upregulated in particular subtypes of cells, as detailed in Supplementary Text 1. Here too we found that Sanity performed best, although sctransform, TPM, and Deconvolution achieved almost equal performance, whereas MAGIC, DCA, and scVI typically performed poorly on this task (Suppl. Fig. 4.27). Supplementary Text 1 provides additional in-depth discussion of Sanity's features and limitations, including its performance for multimodally (Suppl. Fig. 4.28) and very low expressed genes (Suppl. Fig. 4.29), and how the observed absolute expression distributions (Suppl. Fig. 4.30) and sequencing depth determine the accuracy of expression estimates across genes (Suppl. Fig. 4.31).

4.3 Discussion

In this work we developed a new normalization procedure for scRNA-seq data from first principles using only two basic assumptions. First, we characterize a cell's gene expression state by the vector of log transcription quotients (LTQs) across genes,

i.e. the logarithms of the expected fractions of the transcript pool for each gene. Second, to estimate these LTQs from the raw UMI count data, we characterize the prior distribution of LTQs of each gene by just its mean and variance across cells. Given these two assumptions the entire procedure follows from first principles without any tunable parameters, and returns estimated LTQs that correct both for the Poisson noise that is intrinsic to the process of transcription, as well as the sampling noise of the scRNA-seq measurements. Consequently, variation in the inferred LTQs reflect changes in the rates of transcription and mRNA decay of each gene.

Although our procedure makes only minimal assumptions, one may still ask how arbitrary these assumptions are. If one accepts that biological and technical sampling noise do not reflect changes in gene expression state, that expression changes should be measured in terms of fold-changes rather than absolute changes, and that rescaling the expression levels of all genes by a common factor does not change expression state, then LTQs naturally follow as the most general representation of a cell's expression state. Similarly, our prior distribution over LTQs of a gene also aims to minimize the strength of our method's assumptions by using the least assuming, i.e. maximum entropy, distribution consistent with a given mean and variance. Improving on these assumptions would require specific biological information to determine more informative priors on the gene expression states that cells can take on.

Our benchmarking tests indicate that Sanity's normalized expression values outperform those of other methods on basic downstream processing tasks such as clustering cells into subtypes and identifying nearest-neighbor cells. More importantly, we showed that all other methods produce a representation of the data that is distorted in one or more respects.

The simple TPM and closely related Deconvolution methods produce representations of the data that are generally reasonable and perform quite well on downstream tasks such as clustering and identification of differentially expressed genes. The main problem with the TPM method is that the variation in normalized expression levels is dominated by Poisson fluctuations for most genes, and low expressed genes are predicted to be the most variable, whereas in reality these have least evidence of true variability. This also causes the TPM and Deconvolution methods to perform poorly in identifying nearest-neighbor cells, although this can be mitigated to some extent by only considering highly expressed genes. The simple RawCounts method, and the similarly performing scImpute method, suffer from these same problems, but additionally have the problem of not correcting for variation in total UMI count across cells.

The sctransform method outputs z-statistics rather than gene expression estimates. Although this has some advantages, e.g. the method performs well in identifying differentially expressed genes, the clear drawback is that it cannot accurately predict log fold-changes in expression levels, and performs quite poorly

in identifying nearest-neighbor cells.

The sophisticated scVI and DCA methods that use autoencoders to map the data to a low-dimensional manifold perform well on estimating the distances between cells but do this at the cost of strongly distorting the expression levels of individual genes. These methods poorly estimate log fold-changes of genes across cells, produce strong artefactual correlations of the normalized expression values with the total UMI count in each cell, and spuriously predict large numbers of co-expressed genes. Although SAVER performs better in estimating the variances and log fold-changes of genes across cells, it suffers from the same spurious prediction of correlations, as does MAGIC which, in our hands, performed poorly on most tests.

That such spurious correlations are also induced when the TPM normalized expression values are projected onto the top PCA components suggests that they generically result from fitting the data to a lower-dimensional representation. Although it is reasonable to assume that the space of gene expression states that cells take on has much lower dimensionality than the full dimensionality of the transcriptome data, the task of finding such lower-dimensional representations should be clearly distinguished from normalization and noise-correction. Because Poisson sampling noise scales with absolute expression levels, different genes and cells are affected to different extents, and this may be erroneously mistaken for 'structure' in the data. Thus, unless the process of noise removal and normalization is carefully separated from fitting of the data to lower-dimensional representations, artefactual correlations are likely to be introduced.

Finding biologically meaningful lower-dimensional representations of genome-wide gene expression states is one of the most important challenges in the field. However, it is likely a very hard problem in general, and it is unclear to us whether the problem is even solvable with current data. For example, we are not aware of mathematical results that show under what conditions a lower-dimensional manifold embedded in a very high-dimensional space can be reliably reconstructed from a limited number of noisy measurements. We believe that, rather than black box procedures for dimensionality reduction, progress in understanding the genome-wide structure of expression data will crucially depend on connecting transcriptome data to the underlying biophysical mechanisms, e.g. the dynamics of the chromosome, chromatin accessibility at enhancers and promoters, the binding and unbinding of transcription factors, recruitment of the transcription machinery, and the mechanisms of transcription initiation. However, whatever approach is taken to finding lower-dimensional representations of gene expression states, a prerequisite is that the raw data are carefully normalized and corrected for both biological and technical sampling noise. The Sanity method that we presented here aims to provide such normalization methodology.

4.4 Online Methods

4.4.1 A Bayesian method for inferring gene expression states from count data

After motivating how we represent gene expression states of single cells, and to what concrete physical quantities these gene expression states correspond, we introduce our probabilistic model of a scRNA-seq experiment, and calculate the probabilities of the observed raw transcript counts as a function of each cell's expression state. We then explain the Bayesian procedure by which the gene expression states are inferred from the sequencing data, and the outputs that the method provides. Additional discussion of the properties and limitations of Sanity's model are provided in Supplementary Text 1, including a discussion of how Sanity can be used to correct for technical batch effects.

Defining gene expression states

For any given cell c, we want to represent its 'gene expression state' by a vector $\vec{e}_c$, whose components e_{gc} quantify how strongly each gene g is expressed. These gene expression states should satisfy two basic desiderata. First, the gene expression states should have concrete physical interpretation. Second, for each gene g, the difference $e_{gc} - e_{gc'}$ should meaningfully reflect the change in its expression between cells c and c'.

One might think that we could simply take the vector $\vec{m}_c$ of the actual number of mRNAs m_{gc} that exist in cell c for each gene g as the gene expression state of the cell. However, the gene expression process is inherently stochastic due to thermal noise and the low molecule numbers involved, e.g. there are only 1-2 copies of each promoter in a given cell, causing mRNA counts to fluctuate even between cells that are in the same state. To illustrate this, let's imagine a gene that is transcribed at a constant rate λ and whose mRNAs decay at a constant rate μ in every cell. This is as close one can come to having no variation in expression state across cells. However, even in this case the actual number of mRNAs m for this gene will fluctuate across cells according to a Poisson distribution with mean $a = \lambda/\mu$. That is, the probability to find m mRNAs is $P_m = a^m e^{-a}/m!$ which has mean $\langle m \rangle = a$ and variance $\mathrm{var}(m) = a$. Thus, instead of interpreting any change in mRNA number m as a change in gene expression state, it makes more sense to identify changes in gene expression state with changes in the transcription and decay rates λ and μ.

In general, for a given gene g in a given cell, both its transcription rate λ_g and the decay rate μ_g of its mRNAs will vary with time in time t in a potentially complex manner. As illustrated in the cartoon Fig. 4.1A, a large array of different biophysical processes can affect the transcription rate of a given gene including changes in the chromatin state around its locus, the binding and unbinding of

transcription factors (TFs) to its promoter and enhancers, changes in the 3D organization of the chromosome, and so on. Together these processes will determine some time dependent transcription rate $\lambda_g(t)$. Similarly, the rate $\mu_g(t)$ at which mRNAs for gene g decay will depend on the concentrations of RNAses, microRNAs, various RNA binding proteins, and so on. If at some point in time the cell c is sampled and its mRNAs extracted, then the number of mRNAs m_{gc} that one finds for gene g will depend on what the transcription rate $\lambda_g(t)$ and decay rate $\mu_g(t)$ were in the recent past of this cell.

In particular, if we denote the time point at which the cell is sampled as $t = 0$, and denote by $\lambda_g(t)$ and $\mu_g(t)$ the transcription and decay rates a time t into the *past* of the cell, then the expected number of mRNAs $\langle m_{gc} \rangle$ is given by

$$\langle m_{gc} \rangle = \int_0^\infty \lambda_{gc}(t) \exp\left[- \int_0^t \mu_{gc}(\tau) d\tau \right] dt \equiv a_{gc}, \tag{4.1}$$

which we call the 'transcription activity' a_{gc} of gene g in cell c (Fig. 4.1B). Note that a_{gc} is a weighted average of the transcription rates in the past of the cell, where the weights correspond to the probability that an mRNA that was described a time t in the past has survived until now.

Crucially, independent of how $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ have fluctuated in time, the probability to see m_{gc} mRNAs for gene g in cell c is still given by a Poisson distribution with mean a_{gc} [Thattai, 2016] (Fig. 4.1C), i.e.

$$P(m_{gc}|a_{gc}) = \frac{(a_{gc})^{m_{gc}}}{m_{gc}!} e^{-a_{gc}}. \tag{4.2}$$

Thus, independently of how $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ have fluctuated in the cell's past, the number of mRNAs m_{gc} depends on these rates only through the transcription activity a_{gc}. Vice versa, all information about the time-dependent rates $\lambda_g(t)$ and $\mu_g(t)$ that is contained in measurements of mRNA counts in cell c, is contained in the transcription activities a_{gc} for each gene. Thus, we propose to characterize the expression state of a cell by the vector $\vec{a}_c$ of its transcription activities. Note that, as discussed in Supplementary Text 1, it is in principle possible to learn more about the functions $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ by also incorporating information from intronic UMIs of each gene g, e.g. as employed in the RNA velocity approach [La Manno et al., 2018,Bergen et al., 2020]. Although an interesting direction for future extensions of Sanity, we here do not yet incorporate information from intronic UMIs.

Next, we propose that rather than directly representing the gene expression state of the cell by the vector $\vec{a}_c$ of absolute transcription activities a_{gc}, it is beneficial to use the vector $\vec{\alpha}_c$ of *relative* transcription activities, defined as

$$\alpha_{gc} = \frac{a_{gc}}{\sum_{g'} a_{g'c}}, \tag{4.3}$$

which we will refer to as *transcription quotients*, and which correspond to the expected proportions of mRNAs in the cell (Fig. 4.1B). First, it has been shown that, as cell volume increases, cells globally upregulate transcription to maintain approximately constant mRNA concentration [Padovan-Merhar et al., 2015] so that transcriptional activities a_{gc} of all genes are generally expected to scale with cell volume. We argue that a global change in transcriptional activities by a common scale factor S, i.e. $a_{gc} \to Sa_{gc}$ for all genes, does not correspond to a change in gene expression state, but just to a change in cell size. Second, it is well known that, in current scRNA-seq protocols, the rate of capture and sequencing of mRNAs varies significantly across cells [Grün et al., 2014, Stegle et al., 2015] so that there is only a weak quantitative relationship between the total number of sequenced mRNA molecules and the true total mRNA content of a cell. Although it is possible to estimate capture and sequencing efficiencies, at least to some extent, using RNA spike-in controls [Brennecke et al., 2013, Grün et al., 2014], most experiments are performed without such controls. Therefore, for most scRNA-seq datasets it is unclear to what extent variations in total sequenced mRNAs across cells represent biological variability, as opposed to technical variability. Consequently, transcription quotients α_{gc} can generally be much more accurately estimated than absolute transcription activities a_{gc}, because they do not directly depend on capture efficiency. Note that quantifying gene expression by quotients, i.e. transcripts per million transcripts, is also the standard approach in bulk RNA-seq experiments.

Finally, we note that if we were to use differences in transcription quotients of mRNAs $\alpha_{gc} - \alpha_{gc'}$ to quantify the change in expression of gene g between cells c and c', then this change would be proportional to the overall expression level of the gene. That is, a change from 20 to 40 transcripts per million would be considered ten times as large as a change from 2 to 4 transcripts per million. Since the early days of transcriptomics it has been observed [Hoyle et al., 2002] that, as would be expected from the multiplicative effects of fluctuations in rates of various biochemical reactions [Beal, 2017], the relative expression levels of genes in a sample follows a roughly log-normal distribution that covers several orders of magnitude and the variance in absolute expression of a gene across conditions scales with the square of its means expression (Suppl. Fig. 4.7). Consequently, if we were to quantify expression changes directly by the changes $\alpha_{gc} - \alpha_{gc'}$, the expression changes between two cells would be dominated by those of the highest expressed genes. Therefore, it has long become standard to instead use *logarithms* of the expression levels. Indeed, in bulk RNA-seq experiments one also generally finds that the variance in log-expression of a gene across conditions is uncorrelated with its mean expression (Suppl. Fig. 4.7).

Thus, we propose to quantify the gene expression state of a cell by the *logarithms of the transcription quotients* (LTQs) $\log(\alpha_{gc})$ so that an x-fold change in quotient $\alpha_{gc} \to \alpha_{gc'} = x\alpha_{gc}$ corresponds to the same additive change $\log(\alpha_{gc}) \to \log(\alpha_{gc}) + \log(x)$ in LTQ, independent of the absolute value of the quotient α_{gc}.

To define an overall change in expression state between two cells, we still have to combine the changes in LTQ of all genes into a total 'distance'. As motivated in more detail in Supplementary Text 1, we will follow the generally accepted practice of calculating simple Euclidean distances in the space of LTQ vectors, i.e. the squared distance $d^2_{cc'}$ between a pair of cells c and c' is defined as

$$d^2_{cc'} = \sum_g \left[\log(\alpha_{gc}) - \log(\alpha_{gc'})\right]^2 . \tag{4.4}$$

A probabilistic model for a scRNA-seq experiment

The initial steps of scRNA-seq analysis involve basic processing of the raw sequencing reads such as quality control, identification of barcodes to identify the library, the individual cell, and the unique mRNA molecule (if available), and mapping each read to the corresponding genome or transcriptome. The methods used in these steps are similar to methods used for bulk RNA-seq and ChIP-seq and have matured to the point that there is little variability in the results from commonly used tools, e.g. [Love et al., 2015, Bray et al., 2016, Dobin et al., 2013, Genomics, 2020].

The introduction of unique molecule identifiers (UMIs) [Islam et al., 2014] was an important development in scRNA-seq technology in that it avoids PCR amplification noise and allows determining the number of unique mRNA molecules that were captured for each mRNA. It is currently unclear how to realistically model the noise statistics of protocols that do not incorporate UMIs and we will here focus on scRNA-seq protocols that use UMIs.

After basic processing of the raw sequences, the data will consist of a matrix of integers n_{gc} giving the number of captured mRNA molecules for each gene g in each cell c. The key assumption of our probabilistic model is that, in a scRNA-seq experiment, each mRNA molecule in a given cell c has the same probability p_c to be captured and sequenced. This capture probability varies from cell to cell, and has been estimated to be in the range of 10 to 15% [AlJanahi et al., 2018] and up to 30% with the most recent protocols [10X Genomics, 2018]. Under this assumption, the probability of the observed UMI counts n_{gc} in cell c given the transcription quotients α_{gc} is given by a product of Poisson distributions (Fig. 4.1C and Supplementary Methods). Finally, if we marginalize over the unknown capture efficiency p_c we obtain (see Supplementary Methods)

$$P(\vec{n}_c|\vec{\alpha}_c) = \prod_g \left[\frac{(N_c\alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c\alpha_{gc}}\right] , \tag{4.5}$$

where $\vec{n}_c$ is the vector of UMI counts in cell c, $\vec{\alpha}_c$ the vector of transcription quotients in cell c, and N_c the total number of UMIs in cell c (Fig. 4.1D). Crucially, the convolution of the biological Poisson noise and the sampling noise introduced

by the scRNA-seq measurement together still lead to a simple Poisson distribution in terms of the transcription quotients α_{gc} (Supplementary Methods).

Prior probabilities and the Bayesian solution

Having argued that we want to characterize each cell's gene expression state by the vector of LTQs $\log(\alpha_{gc})$, and having determined how likely it is to observe UMI counts $\vec{n}_c$ given the transcription quotients $\vec{\alpha}_c$, i.e. equation (4.5), we now want to invert this relation and estimate the LTQs from the observed UMI counts. The uniquely consistent set of mathematical procedures for doing this is generally referred to as Bayesian probability theory [Jaynes, 2003].

This calculation requires that we specify a *prior* probability distribution that represents the prior information we want to assume about how LTQs may vary across cells, before obtaining the expression data. As we aim to minimize the number of assumptions, our model will not assume any dependence structure between the LTQs of different genes, i.e. we will not assume *a priori* that the gene expression data derives from a low-dimensional manifold. We will also not assume that the LTQs of gene across cells follow a particular distribution. The only thing that we will assume is that, for each gene, the prior distribution of LTQs $\log(\alpha_{gc})$ can be characterized by its mean μ_g and variance v_g.

Without loss of generality, we rewrite the transcription quotients α_{gc} in terms of an average log-quotient μ_g and cell-specific log fold-changes δ_{gc}, i.e. $\alpha_{gc} = e^{\mu_g + \delta_{gc}}$. With this reparametrization, the δ_{gc} derive from a prior probability distribution with mean zero and variance v_g. Given that we only specify the variance of the distribution of the δ_{gc} to be v_g, this implies that the prior corresponds to the maximum entropy distribution consistent with this constraint, which is a Gaussian distribution [Jaynes, 2003]. Importantly, this does not mean that we assume that the log fold-changes δ_{gc} follow a Gaussian distribution. Indeed, as we demonstrate in Supplementary Text 1, the δ_{gc} that our method infers upon seeing the data do not necessarily follow Gaussian distributions. For example, if a gene is bimodally distributed, the method correctly infers this in general (Suppl. Fig. 4.28).

In the Supplementary Methods we show in detail how this model can be solved to estimate, for each gene g:

1. The mean LTQ μ_g and its error-bar $\delta\mu_g$.
2. The estimated variance v_g of the changes in LTQs δ_{gc} across cells.
3. For each cell c, the estimated log fold-changes δ^*_{gc} and an error-bar ϵ_{gc} on each of these.

Note that the δ^*_{gc} provide estimates for how much the transcription and decay rates of each gene g in cell c differ from their average rates, and thus correct for both the intrinsic biological Poisson fluctuations as well as the finite sampling fluctuations inherent in the scRNA-seq measurements.

4.4.2 Other methods for scRNA-seq normalization

To assess the performance of Sanity we compare it with a number of other methods for normalization/imputation from scRNA-seq data. Here we introduce these other methods and highlight the ways in which their approaches differ from Sanity's. Apart from tools from the recent literature, we include two basic normalization procedures that are widely used. First, the simplest approach to estimating gene expression levels e_{gc} from scRNA-seq data is to simply log-transform the observed number of UMIs n_{gc} after adding a *pseudocount* p to avoid problems with zero counts $n_{gc} = 0$, i.e.

$$e_{gc} = \log(n_{gc} + p). \tag{4.6}$$

A typical choice for the pseudo-count is $p = 1$, because it attenuates fluctuations in n_{gc} on the order of magnitude corresponding to the resolution of the experimental measurements. We refer to this normalization, with $p = 1$, as the *RawCounts* normalization, since it essentially just log-transforms the raw UMI counts.

However, the total number N_c of mRNAs captured and sequenced from an individual cell c can vary substantially due to fluctuations in capture efficiency and sequencing depth, as well as differences in cell size. Consequently, the RawCounts procedure introduces systematic correlations between the expression levels e_{gc} and the total number of UMIs N_c that were sequenced from cell c. Thus, the most commonly used normalization approach is to first divide the RawCounts n_{gc} by the total count N_c and then multiply by a typical total count N before adding a pseudocount and log transforming, i.e.:

$$e_{gc} = \log\left[\frac{n_{gc}}{N_c}N + 1\right]. \tag{4.7}$$

Here we take for the typical total count N the median of the counts N_c across all cells. In a slight abuse of terminology, we will call this normalization the *TPM* normalization because of its close connection to the transcripts per million normalization used in bulk RNA-seq (which corresponds to setting $N = 10^6$).

Given the definition of the LTQs as logarithms of relative expression levels, a reader may wonder how our approach is even different from this standard TPM procedure. Indeed, for a cell with total count N_c and LTQs $\log(\alpha_{gc}) = \mu_g + \delta_{gc}$, the expected number of UMI is

$$\langle n_{gc} \rangle = N_c e^{\mu_g + \delta_{gc}}, \tag{4.8}$$

which might suggest that if we simply divide n_{gc} by N_c and log-transform the result, we would end up with the LTQ $\mu_g + \delta_{gc}$. However, the actual UMI counts n_{gc} are not the same as the expectations $\langle n_{gc} \rangle$. That is, the n_{gc} are measured quantities that contain Poisson noise both due to the intrinsic stochasticity of gene expression and due to the measurement process. Importantly, instead of

n_{gc} differing from $\langle n_{gc} \rangle$ by noise of a constant size, the size of the Poisson noise depends on the expected count $\langle n_{gc} \rangle$ itself. In addition, since UMI counts n_{gc} are very small for most genes, the noise is typically larger than the true variation in LTQ across cells. Therefore, to estimate the LTQ of each gene in each cell, it is crucial to account for this Poisson noise and this is one of Sanity's main aims.

Beyond the simple RawCount and TPM normalization methods, we compare Sanity's performance with those of the following recently published tools:

1. DCA [Eraslan et al., 2019], which uses a deep learning based autoenconder together with a zero-inflated negative binomial noise model.
2. Deconvolution [Lun et al., 2016], which is similar to the TPM method, but uses a more sophisticated approach to normalize for the variation in sequencing depth across cells.
3. MAGIC [van Dijk et al., 2018], which uses diffusion of measured gene expression states between cells with similar expression profiles.
4. SAVER [Huang et al., 2018], which assumes negative binomial counts distributions n_{gc} and models the underlying rates using Poisson LASSO regression with the expression levels of other genes.
5. scImpute [Li and Li, 2018], which focuses mainly on correcting 'dropouts', i.e. data points for which $n_{gc} = 0$.
6. sctransform [Hafemeister and Satija, 2019], which uses regularized negative binomial regression and reports Pearson residuals of this regression rather than estimated expression values.
7. scVI [Lopez et al., 2018], which uses a deep neural network based autoencoder together with a zero-inflated negative binomial noise model.

Note that, with the exception of RawCounts and scImpute, all these methods seek to normalize the expression levels for the total UMI count per cell. In contrast to Sanity, RawCounts, TPM, and Deconvolution, all other methods seek to remove noise by fitting the data to lower dimensional representations. Specifically, in SAVER and sctransform the parameters of each gene's negative binomial model are fitted by using information from other genes, in scImpute zero values are corrected for by using information from neighboring cells, in MAGIC the entire expression profile of each cell is estimated using information of neighboring cells, and in DCA and scVI the autoencoders effectively force a lower dimensional representation of the distribution of cells in gene expression space.

Many of the models above use a negative binomial or zero-inflated negative binomial to model the distribution of UMI counts of a gene across cells and the reader may wonder how Sanity's noise model relates to these models. In Supplementary Text 1 we explain why, as discussed recently [Svensson, 2020], no zero

inflation is necessary and discuss the relationship of Sanity's model with negative binomial noise models.

We used default parameters for all methods except for scVI, where we adapted settings based on direct feedback of the developers of scVI (the default parameter `n_epochs=20` was increased to 400 and we used the recently added `get_sample_scale` instead of `imputation` method to get predicted expression values).

Since all methods report expression values in linear space, we log-transformed all expression values. MAGIC sometimes reports 0 or even negative values and, as suggested by its developers, we first set all negative values to 0 and then add a pseudocount of 1 to all expression values (including the nonzero ones) before log-transforming. Similarly, scImpute reports some zero values and we added a pseudocount of 1 to all the expression values.

Directly comparing the results of sctransform with those of the other methods is complicated by the fact that, in contrast to all other methods, sctransform does not provide estimated gene expression values, but z-statistics z_{gc} that quantify how significantly the expression of gene g in cell c deviates from what would be expected from the negative binomial model. The authors of the sctransform paper suggest that these z-statistics should be used for downstream analyses. Since the z-statistics are variance normalized and centered around zero, we use the z-statistic z_{gc} equivalently to the log-fold changes δ_{gc}. Finally, in the negative binomial fit, sctransform fits the expected mean log-expression μ_{gc} of gene g in cell c to a function of the form $\mu_{gc} = \beta_0 + \beta_1 \log(N_c)$, with N_c the total UMI count of cell c. To calculate a predicted average expression for gene g we use $\mu_g = \beta_0 + \beta_1 \log(N)$, with N the median total UMI count.

4.4.3 Test datasets

To assess the performance of the different methods we used a collection of datasets for which annotation of the sequenced cell types was available. These were (labelled by the first author of the publication):

1. *Grün*: 160 mouse embryonic stem cells and 160 corresponding aliquots consisting of 80 cells from culture in 2i medium, 80 cells from culture in serum, and 80 aliquots for each condition that were created by pooling RNA from the cells, and then splitting the pool into single-cell mRNA equivalents [Grün et al., 2014].

2. *Zeisel*: 3'005 cells from the somatosensory cortex and from the CA1 region of the mouse hippocampus, annotated into 7 cell types [Zeisel et al., 2015].

3. *Baron*: 1'937 human pancreatic cells annotated into 14 cell types [Baron et al., 2016].

4. *Chen*: 14,437 adult mouse hypothalamus cells annotated into 15 clusters [Chen et al., 2017].

5. Three datasets from *LaManno* [La Manno et al., 2016]:
   (a) *LaManno/Embryo*: 1'977 ventral mid-brain cells from human embryo annotated into 25 classes.
   (b) *LaManno/ES*: 1'715 human embryonic stem cells annotated into 17 classes.
   (c) *LaManno/MouseEmbryo*: 1'907 ventral mid-brain cells from mouse embryo annotated into 26 classes.

In addition to these real datasets we also constructed two simulated datasets as detailed in the Supplementary Methods. The distributions of means and variances in log-expression, as well as the distribution of total UMI count per cell were chosen so as to mimic the statistics of an arbitrarily chosen real dataset, for which we chose the *Baron* dataset (see Fig. 4.9). In the first simulated dataset the expression profiles of all genes were drawn randomly and independently, so that there are no expression correlations by construction. This dataset we used to test the ability of different methods to correctly estimate true means, variances, and log fold-changes in expression of each gene, and to assess the extent to which different methods spuriously predicted co-expression of genes. The second simulated dataset was constructed by performing a branched random walk in the high-dimensional gene expression space so that the true expression profiles of the cells fall on a tree. We used this dataset to test the ability of different methods to identify the k nearest-neighbor cells of each cell.

Data Availability

The raw UMI count tables for each of the scRNA-seq datasets, as well as all the normalized expression values as inferred by each of the methods are freely available from `https://doi.org/10.5281/zenodo.4009187`.

Code Availability

Sanity was implemented in C and is freely available for download at `https://github.com/jmbreda/Sanity`. Besides Sanity itself, we also provide code for estimating pairwise distances between cells. In addition, at the same github side we provide a collection of scripts and supplementary files that should allow other researchers to reproduce the results presented in this publication.

Acknowledgements

This work was supported by the Swiss National Science Foundation grant No. 310030_184937. Calculations were performed at sciCORE (http://scicore.unibas.ch/), the scientific computing core facility of the University of Basel.

Author Contributions

E.v.N. developed the theoretical formalism. J.B. and E.v.N. developed the implementation, and designed the benchmarking. J.B. performed all computations, analyses, and simulations. J.B., M.Z., and E.v.N. interpreted the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

4.5 Supplementary Methods

4.5.1 Sanity

We denote, for each cell c and each gene g, the transcription rate a time t in the past as $\lambda_{gc}(t)$ and the decay rate of its mRNAs a time t in the past as $\mu_{gc}(t)$. Given these time-dependent transcription and decay rates, we define the *transcription activity* a_{gc} of gene g in cell c as the expected number of mRNAs $\langle m_{gc} \rangle$ which can be written as the following integral

$$a_{gc} = \langle m_{gc} \rangle = \int_0^\infty dt \lambda_{gc}(t) \exp\left[- \int_0^t \mu_{gc}(s) ds \right]. \tag{4.9}$$

That is, the transcription activity a_{gc} is a weighted time average of the recent transcription rates a time t in the past, with the weight equal to the expected fraction of the mRNAs produced a time t ago that survive until now.

Conditioned on the transcription activity a_{gc}, the distribution of the actual number of mRNAs m_{gc} for gene g in cell c is given by a simple Poisson distribution

$$P(m_{gc}|a_{gc}) = \frac{(a_{gc})^{m_{gc}}}{m_{gc}!} e^{-a_{gc}}. \tag{4.10}$$

We now assume that, in the scRNA-seq measurement, each mRNA existing in cell c has a probability p_c to be captured and sequenced. Given this, the probability that precisely n_{gc} unique mRNAs will be sequenced for gene g in cell c is given by

$$P(n_{gc}|a_{gc}, p_c) = \sum_{m_{gc}=n_{gc}}^{\infty} \binom{m_{gc}}{n_{gc}} (p_c)^{n_{gc}} (1-p_c)^{m_{gc}-n_{gc}} P(m_{gc}|a_{gc}) \tag{4.11}$$

$$= \frac{(p_c a_{gc})^{n_{gc}}}{n_{gc}!} e^{-p_c a_{gc}}, \tag{4.12}$$

which is still a Poisson distribution.

Next, we define the transcription activity a_{gc} as a product of the total transcription activity $A_c = \sum_g a_{gc}$ in cell c and a *transcription quotient* α_{gc}:

$$\alpha_{gc} = \frac{a_{gc}}{\sum_{g'} a_{g'c}} = \frac{a_{gc}}{A_c}. \tag{4.13}$$

Note that α_{gc} is the expected *fraction* of all mRNAs in cell c that are mRNAs for gene g, i.e. we have

$$\langle \frac{n_{gc}}{\sum_{g'} n_{g'c}} \rangle = \alpha_{gc}. \tag{4.14}$$

If we define the cell dependent constant $\lambda_c = p_c A_c$, then we can rewrite this Poisson distribution as

$$P(n_{gc}|\alpha_{gc}, \lambda_c) = \frac{(\lambda_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-\lambda_c \alpha_{gc}}, \tag{4.15}$$

and the probability for the entire data-set in cell c has the form:

$$P(\vec{n}_c|\vec{\alpha}_c, \lambda_c) = \prod_g \left[\frac{(\lambda_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-\lambda_c \alpha_{gc}} \right], \tag{4.16}$$

where the notation $\vec{n}_c$ refers to all counts n_{gc} for cell c, and $\vec{\alpha}_c$ refers to all transcription quotients α_{gc} for cell c.

In the next step we remove the dependence on the cell-dependent scale factor λ_c. If one integrates over λ_c (using a scale prior $1/\lambda_c$) one finds that the probability of the data for one cell is a simple multinomial in the transcription quotients α_{gc}, i.e

$$P(\vec{n}_c|\vec{\alpha}_c) = N_c! \prod_g \frac{(\alpha_{gc})^{n_{gc}}}{n_{gc}!}. \tag{4.17}$$

Note that in this equation the α_{gc} of different genes are subtly coupled together through the constraint $\sum_g \alpha_{gc} = 1$ and this makes further analytical treatment difficult. We will thus make the key approximation that, for each gene g, the fluctuations of α_{gc} across cells can be treated as independent of the fluctuations of the α_{gc} of the other genes. This approximation is accurate as long as the total transcription activity A_c is spread over many genes and not dominated by one or a few very highly expressed genes.

We first formally decouple the genes by, for each gene g, integrating over all the $\alpha_{g'c}$ of all other genes g'. The multinomial distribution over all genes then becomes a binomial for a single gene g:

$$P(n_{gc}, N_c|\alpha_{gc}) = \binom{N_c}{n_{gc}} (\alpha_{gc})^{n_{gc}} (1 - \alpha_{gc})^{N_c - n_{gc}}. \tag{4.18}$$

Whenever $\alpha_{gc} \ll 1$ and N_c is large, this binomial is well approximated by a Poisson with mean $N_c \alpha_{gc}$. We then obtain for the probability of the observed counts $\vec{n}_c$ in

cell c:

$$P(\vec{n}_c|\vec{\alpha}_c) = \prod_g \left[\frac{(N_c\alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c\alpha_{gc}} \right]. \tag{4.19}$$

That is, the number of sequenced mRNAs n_{gc} for each each gene g in cell c is still a Poisson distribution with expectation value $N_c\alpha_{gc}$. The probability of the entire dataset of counts $\{n\}$ given all transcription quotients $\{\alpha\}$ is given by simply taking the product of this expression over all cells, i.e

$$P(\{n\}|\{\alpha\}) = \prod_{c,g} \left[\frac{(N_c\alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c\alpha_{gc}} \right]. \tag{4.20}$$

Instead of trying to estimate the α_{gc} for all genes at once, we will focus on one specific gene g at a time, and infer how α_{gc} varies across the cells c. Note that if we collect all the terms that depend on the α_{gc} of single gene g we obtain

$$P(\vec{n}_g|\vec{\alpha}_g) = \prod_c \left[\frac{(N_c\alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c\alpha_{gc}} \right], \tag{4.21}$$

where $\vec{n}_g$ is the vector of counts for gene g and $\vec{\alpha}_g$ is the vector of transcription quotients for gene g.

Finally, without loss of generality, we will write the transcription quotients α_{gc} in terms of the average quotient of the gene α_g and a log-fold change δ_{gc} in a given cell c, i.e.

$$\alpha_{gc} = \alpha_g e^{\delta_{gc}}. \tag{4.22}$$

In terms of these parameters we have

$$P(\vec{n}_g|\alpha_g, \vec{\delta}_g) = \left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!} \right) \alpha_g^{n_g} \exp\left[\sum_c n_{gc}\delta_{gc} - \alpha_g \sum_c N_c e^{\delta_{gc}} \right], \tag{4.23}$$

where n_g is the total number of sequenced mRNAs for gene g.

Marginalizing over the average transcription quotient α_g

We now first focus on estimating the log fold-changes δ_{gc}. We return to estimating the overall average transcription quotient α_g once we have determined these. To marginalize expression (4.23) over α_g we use a simple uniform prior $P(\alpha_g)d\alpha_g \propto d\alpha_g$. Integrating with this uniform prior from 0 to ∞ we obtain

$$P(\{n_g\}|\{\delta_g\}) = \left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!} \right) \Gamma(n_g+1) \exp\left(\sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left[\sum_c N_c e^{\delta_{gc}} \right] \right). \tag{4.24}$$

Note that, because α_g is a fraction, we should have really only integrated from 0 to 1, but as long as each gene is only responsible for a small fraction of all UMIs in the cell, the only contribution to the integral comes from values of α_g much smaller than 1, and we can extend the range of the integral to infinity without loss of accuracy. That is, we approximate the incomplete gamma function integral with a complete gamma function integral. To assess the accuracy of this approximation, the key parameters are the total number n_g of UMIs for gene g and the quantity $N_g = \sum_c N_c e^{\delta_{gc}}$. Since $e^{\delta_{gc}}$ is order 1 on average, N_g is on the order of the total number of UMIs in the entire dataset, i.e. $N_g \approx \sum_c N_c = N$. Let $f_g = n_g/N$ denote the fraction of all UMIs in the dataset that are for gene g. Approximating the incomplete gamma function with a complete gamma function then leads to a relative error of order $N^{-N(1-f_g)}$, which is very small as long as f_g is not close to 1. That is, the approximation will be very accurate as long as no gene is responsible for almost all of the UMIs in the data. This always holds in practice.

Note also that the factor $\left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!}\right)\Gamma(n_g+1)$ is determined entirely by the counts and does not depend on the δ_{gc}, and we will neglect this prefactor from now on.

Including prior probabilities for the δ_{gc}

We next introduce prior probabilities over the log fold-changes δ_{gc}. Assuming only that the δ_{gc} for gene g have a variance v_g and mean zero, we use the maximum entropy distribution consistent with these constraints, which is a Gaussian

$$P(\delta_{gc}|v_g) = \frac{1}{\sqrt{2\pi v_g}} \exp\left[-\frac{\delta_{gc}^2}{2v_g}\right]. \tag{4.25}$$

Thus, the prior over the vector $\vec{\delta}_g$ of log fold-changes for the C cells is given by

$$P(\vec{\delta}_g|v_g) \propto (v_g)^{-C/2} \exp\left(-\frac{1}{2v_g}\sum_{c=1}^{C}\delta_{gc}^2\right). \tag{4.26}$$

Combining the prior with the likelihood we obtain

$$P(\vec{n}_g, \vec{\delta}_g|v_g) = (v_g)^{-C/2} \exp\left(-\frac{1}{2v_g}\sum_c \delta_{gc}^2 + \sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left[\sum_c N_c e^{\delta_{gc}}\right]\right), \tag{4.27}$$

up to a prefactor that does not depend on the parameters δ_{gc} and v_g.

Calculating $P(\vec{n}_g|v_g)$ using the Laplace approximation

We next focus on calculating the probability $P(\vec{n}_g|v_g)$ of the data given only the variance v_g. To obtain the probability $P(\vec{n}_g|v_g)$, we need to integrate over all

possible δ_{gc}. As the integral is close to Gaussian in form, we will assume we can approximate the integral by the Laplace approximation, i.e. by approximating the log-likelihood $L(\vec{\delta}_g, v_g) = \log\left[P(\vec{n}_g, \vec{\delta}_g|v_g)\right]$ by expanding it to second order around its maximum. The log-likelihood has the form

$$L(\vec{\delta}_g, v_g) = -\frac{C}{2}\log(v_g) - \frac{1}{2v_g}\sum_c \delta_{gc}^2 + \sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left(\sum_c N_c e^{\delta_{gc}}\right). \tag{4.28}$$

Taking derivatives with respect to the δ_{gc}, the equations for the optimum become

$$-\frac{\delta_{gc}}{v_g} + n_{gc} - (n_g+1)\frac{N_c e^{\delta_{gc}}}{\sum_{\tilde{c}} N_{\tilde{c}} e^{\delta_{g\tilde{c}}}} = 0\ \forall c. \tag{4.29}$$

To solve this equation we are going to multiply the equation by v_g and then define the c-independent quantity

$$e^{q_g} = \sum_c N_c e^{\delta_{gc}}, \tag{4.30}$$

the normalized quantities

$$f_{gc} = e^{-q_g} N_c e^{\delta_{gc}}, \tag{4.31}$$

which sum to 1, i.e. $\sum_c f_{gc} = 1$, and the c-dependent quantities

$$y_{gc} = v_g n_{gc} + \log(N_c), \tag{4.32}$$

which are directly determined by v_g and the data.

In terms of these quantities the equations for the optimum become

$$\log(f_{gc}) + v_g(n_g+1)f_{gc} = -q_g + y_{gc}\ \forall c, \tag{4.33}$$

whose solution is

$$f_{gc} = \frac{W\left[e^{-q_g+y_{gc}} v_g(n_g+1)\right]}{v_g(n_g+1)}, \tag{4.34}$$

with $W(x)$ the Lambert W-function (also called productlog). Note, however, that the solution depends on q_g, which itself depends on the f_{gc}. However, since $\sum_c f_{gc} = 1$ per definition, we can sum equation (4.34) over c to obtain the following consistency equation for q_g

$$\sum_c \frac{W\left[e^{-q_g+y_{gc}} v_g(n_g+1)\right]}{v_g(n_g+1)} = 1. \tag{4.35}$$

In the above equation, everything is determined either by the data (n_{gc}, n_g, and N_c) or the variance v_g, except for the unknown constant q_g, which needs to be solved for numerically. We can perform a binary search to find the value of q_g for

which equation (4.35) is satisfied. Note also that the expression on the left hand side of equation (4.35) is a monotonically decreasing function of q_g, guaranteeing that there is only a single solution for q_g.

Once q_g has been determined, we obtain the f_{gc} from equation (4.34) and we obtain the optimal δ^*_{gc} as

$$\delta^*_{gc} = \log(f_{gc}) - \log(N_c) + q_g. \tag{4.36}$$

Note that these optimal δ^*_{gc} are functions of the variance v_g, which we from now on will express explicitly in our notation.

Substituting the optimal $\delta^*_{gc}(v_g)$ into equation (4.28) we obtain the optimal log-likelihood $L_*(v_g)$. By expanding the log-likelihood to second order around its maximum, the probability $P(\vec{n}_g, \vec{\delta}_g|v_g)$ can then be rewritten as

$$P(\vec{n}_g, \vec{\delta}_g|v_g) = \exp\left[L_*(v_g) - \frac{1}{2}\sum_{c,\tilde{c}}(\delta_{gc} - \delta^*_{gc}(v_g))M^g_{c\tilde{c}}(\delta_{g\tilde{c}} - \delta^*_{g\tilde{c}}(v_g))\right], \tag{4.37}$$

where the matrix M^g is given by the second derivatives of the log-likelihood around its optimum, i.e.

$$\frac{\partial^2 L}{\partial\delta_{gc}\partial\delta_{g\tilde{c}}}\|_* = -M^g_{c\tilde{c}}. \tag{4.38}$$

We find

$$M^g_{c\tilde{c}} = \left((n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}\right)\delta_{c\tilde{c}} - (n_g+1)f^*_{gc}(v_g)f^*_{g\tilde{c}}(v_g). \tag{4.39}$$

The integral over the likelihood can now be easily written in terms of the determinant of the matrix M^g, giving us for the marginal probability of the data as a function of v_g:

$$P(\vec{n}_g|v_g) = \int P(\vec{n}_g, \vec{\delta}_g|v_g)d\vec{\delta}_g = \frac{e^{L_*(v_g)}}{\sqrt{\det(M^g)}}. \tag{4.40}$$

Finally, given the relatively simple structure of the matrix M^g, we use the *matrix determinant lemma* toq write the determinant as

$$\det(M^g) = \left(1 - \sum_c \frac{(n_g+1)(f^*_{gc}(v_g))^2}{(n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}}\right)\prod_c\left((n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}\right). \tag{4.41}$$

Posterior $P(v_g|\vec{n}_g)$ over variance v_g

To obtain a posterior over the variance v_g we need a prior over the variance v_g, for which we will use a scale prior, i.e. uniform in the logarithm of v_g:

$P(v_g)dv_g \propto d\log(v_g)$. Note, however, that our solution of $P(\vec{n}_g|v_g)$ involved a numerical determination of q_g, so that we do not have an analytical formula for $P(v_g|\vec{n}_g)$. In order to approximate the full posterior $P(v_g|\vec{n}_g)$ we pick a range $[v_{\min}, v_{\max}]$ within which we presume all v_g fall, divide this range into B bins of equal size in $\log(v_g)$, and calculate $P(\vec{n}_g|v_g)$ for each bin b. Per default we choose $[v_{\min}, v_{\max}] = [0.001, 50]$ since this covers the range of observed variances in the datasets we considered. Trading off speed versus accuracy we chose $B = 160$ bins by default, so that the variance increase by a little under 7% from one bin to the next. However, if desired these values can be changed by the user.

Let v_b denote the variance of bin b and L_b the log-likelihood $\log[P(\vec{n}_g|v_b)]$. We then approximate the full posterior $P(v_b|\vec{n}_g)$ by a distribution over a finite number of points:

$$P(v_b|\vec{n}_g) = \frac{e^{L_b}}{\sum_{b'=1}^{B} e^{L_{b'}}}. \tag{4.42}$$

We note that there are of course much more sophisticated ways of performing this numerical integral. However, as becomes clear below, we need to perform integrals over v_g with weights $P(v_g|\vec{n}_g)$ many times, i.e. for each δ_{gc} and its error bar ϵ_{gc}. Therefore, it is efficient to calculate a fixed set of weights e^{L_b} for binned values v_b once, and then approximate all integrals by summing over the same bins with the same vector of weights. In any case, we have explicitly checked that, for the datasets analyzed here, increasing the number of bins does not alter any of the results.

The posterior $P(\vec{\delta}_g|\vec{n}_g, v_g)$ of log-fold changes given a variance v_g

For a given value of the variance v_g, the posterior distribution over the log fold-changes δ_{gc} is given by a multi-variate Gaussian with means $\langle\delta_{gc}\rangle = \delta^*_{gc}(v_g)$ and a covariance matrix C given by the inverse of the matrix M^g. In particular, the variances $\mathrm{var}(\delta_{gc})$ of the log fold-changes across cells are given by the diagonal elements of the inverse of M^g. Fortunately, given the relatively simple structure of the matrix M^g, we can also obtain analytical expressions for these variances. In particular, the components (c, c) of the inverse of M^g are given by the ratio of the minor $[M^g]_{(c,c)}$ (the determinant of matrix M^g with the cth row and column

removed) and the determinant of the full matrix. We have

$$\begin{aligned}
\mathrm{var}(\delta_{g\tilde{c}}) &= \frac{[M^g]_{\tilde{c},\tilde{c}}}{\det(M^g)} && (4.43)\\
&= \frac{\left(1-\sum_{c\neq\tilde{c}}\frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right)\prod_{c\neq\tilde{c}}\left((n_g+1)f^*_{gc}+\frac{1}{v_g}\right)}{\left(1-\sum_{c}\frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right)\prod_{c}\left((n_g+1)f^*_{gc}+\frac{1}{v_g}\right)} && (4.44)\\
&= \frac{\left(1-\sum_{c\neq\tilde{c}}\frac{(n_g+a){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right)}{\left(1-\sum_{c}\frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right)\left((n_g+1)f^*_{gc}+\frac{1}{v_g}\right)}, && (4.45)
\end{aligned}$$

where again it should be noted that the f^*_{gc} are themselves functions of v_g.

A technical complication arises in estimating the variance $\mathrm{var}(\delta_{gc})$ when the observed number of UMIs is zero. That is, when $n_{gc}=0$ the log-likelihood $L(\vec{\delta}_g, v_g)$ can be a highly asymmetric function of δ_{gc} around its maximum $\delta^*_{gc}(v_g)$. In particular, whereas the fact that no UMIs were observed, i.e. $n_{gc}=0$, ensures that the log-likelihood decreases quickly as δ_{gc} increases above $\delta^*_{gc}(v_g)$, it drops only slowly with decreasing δ_{gc}. That is, when no UMI are observed, we can give a reasonably tight upper bound on δ_{gc}, but $n_{gc}=0$ is consistent with arbitrarily low δ_{gc}. Since we want to summarize the accuracy of the estimate $\delta^*_{gc}(v_g)$ with a single (i.e. symmetric) error bar we need to resolve this asymmetry and we choose to set $\mathrm{var}(\delta_{gc})$ from its upper bound for cases with $n_{gc}=0$. In particular, note that for a Gaussian distribution with mean μ and variance σ^2, the difference between the log-likelihood at the optimum μ and at $\mu+\sigma$ is $L(\mu)-L(\mu+\sigma)=(\mu+\sigma-\mu)^2/(2\sigma^2)=1/2$. We thus define the $\sigma_{gc}=\sqrt{\mathrm{var}(\delta_{gc})}$ such that the difference between the log-likelihood at δ^*_{gc} and $\delta^*_{gc}+\sigma_{gc}$ is $1/2$, i.e. the solution of

$$L(\delta^*_{gc})-L(\delta^*_{gc}+\sigma_{gc})=\frac{\sigma_{gc}(2\delta^*_{gc}+\sigma_{gc})}{2v_g}+(n_g+1)\log(1+f^*_{gc}(e^{\sigma_{gc}}-1))=\frac{1}{2}, \quad (4.46)$$

which we determine numerically.

Final estimates $\langle\delta_{gc}\rangle$ and error-bars ϵ_{gc}

For each value of v_g, we have determined the posterior probability $P(v_g|\vec{n}_g)$, and given a variance v_g we have a Gaussian posterior distribution $P(\vec{\delta}_g|\vec{n}_g, v_g)$ over the log fold-changes, with means $\delta^*_{gc}(v_g)$ and variances $\mathrm{var}(\delta_{gc})(v_g)$. Using these, we can now calculate final estimates of the log fold-changes δ_{gc}. In particular, the expectation value $\langle\delta_{gc}\rangle$ is given by the integral

$$\langle\delta_{gc}\rangle=\int dv_g d\vec{\delta}_g\delta_{gc}P(\vec{\delta}_g|\vec{n}_g,v_g)P(v_g|\vec{n}_g)=\int dv_g\delta^*_{gc}(v_g)P(v_g|\vec{n}_g), \quad (4.47)$$

which we calculate numerically with the same weighted sum over the B bins.

Similarly, we find for the overall error-bar ϵ_{gc}^2

$$\begin{aligned}\epsilon_{gc}^2 &= \langle(\delta_{gc})^2\rangle - \langle\delta_{gc}\rangle^2 && (4.48)\\ &= \int dv_g \left[\mathrm{var}(\delta_{gc})(v_g) + \left(\delta_{gc}^*(v_g)\right)^2\right] P(v_g|\vec{n}_g) - \langle\delta_{gc}\rangle^2 && (4.49)\\ &= \int dv_g \left(\mathrm{var}(\delta_{gc})(v_g) + \left(\delta_{gc}^*(v_g) - \langle\delta_{gc}\rangle\right)^2\right) P(v_g|\vec{n}_g), && (4.50)\end{aligned}$$

which we also calculate numerically using the same weighted sum over the B bins.

Sanity returns, for each gene g in each cell c, both the estimated log fold-change $\langle\delta_{gc}\rangle$ and its error-bar ϵ_{gc}.

Mean expression $\langle\log(\alpha_g)\rangle$

Once we have fitted a set of $\delta_{gc}^*(v_g)$ for each v_g, and determined the posterior $P(v_g|\vec{n}_g)$ we can now easily estimate the mean log quotient $\mu_g = \log(\alpha_g)$ of each gene. Returning to equation (4.23), and marginalizing over the δ_{gc} using the Laplace approximation, we find that the posterior over α_g is proportional to the expression (4.23) in which the δ_{gc} have been set to $\delta_{gc}^*(v_g)$:

$$P(\bar{\alpha}_g|\vec{n}_g, v_g) \propto (\alpha_g)^{n_g} \exp\left[-\alpha_g e^{q_g(v_g)}\right], \tag{4.51}$$

where n_g is the total number of UMIs captured for gene g, $e^{q_g(v_g)} = \sum_c N_c e^{\delta_{gc}^*(v_g)}$ as defined above, and we have explicitly indicated that q_g is a function of the variance v_g.

Using (4.51) the expectation value of $\log(\alpha_g)$ at a given value of the variance v_g is given by

$$\langle\log(\alpha_g)\rangle_{v_g} = \psi(n_g + 1) - q_g(v_g), \tag{4.52}$$

where $\psi(x)$ is the digamma function, i.e. the derivative of the logarithm of the gamma function. Note also that, since n_g is an integer, we have $\psi(n_g+1)$ is simply related to the Harmonic numbers, i.e. $\psi(n_g+1) = -\gamma + \sum_{k=1}^{n_g} 1/k$, with $\gamma \approx 0.577$ the Euler–Mascheroni constant.

To get a final estimate $\mu_g = \langle\log(\alpha_g)\rangle$ we obtain the weighted average over the variance v_g, i.e.

$$\mu_g = \psi(n_g + 1) - \int dv_g q_g(v_g) P(v_g|\vec{n}_g) = \psi(n_g + 1) - \langle q_g\rangle. \tag{4.53}$$

Error bar on mean expression

Going back to equation (4.51) we find that the variance in $\log(\alpha_g)$, at a given value of the variance v_g, is given by the derivative of the digamma function:

$$\mathrm{var}(\log(\alpha_g))_{v_g} = \psi_1(n_g + 1), \tag{4.54}$$

with $\psi_1(x)$ the derivative of the digamma function, which is also called the trigamma function. Note that this variance is independent of v_g.

The final error-bar $\delta\mu_g$ for $\log(\alpha_g + 1)$ is then given by

$$(\delta\mu_g)^2 = \psi_1(n_g + 1) + \int dv_g \left(q_g(v_g) - \langle q_g \rangle\right)^2 P(v_g|\vec{n}_g). \tag{4.55}$$

Note that, as for the calculation of the log fold-changes, these integrals over v_g are approximated by sums over the same set of B bins.

4.5.2 Simulated datasets

We created two simulated datasets. In the first dataset the expression patterns of all genes were chosen randomly and independently so that there are no correlations between the gene expression patterns of different genes. In the second dataset we created expression profiles by performing a branched random walk in gene expression space as described below. For both simulated datasets we matched the gene expression statistics to those of the Baron et al. dataset [Baron et al., 2016]. The first dataset was created as follows:

1. The dataset contained $N_{\text{gene}} = 16'016$ genes and $N_{\text{cell}} = 1'937$ cells.
2. For each gene in the Baron et al. dataset, we calculated the total number of UMI N_g in the data and defined its mean LTQ as $\mu_g = \log(N_g / \sum_{g'} N_{g'})$. In the simulated dataset, each gene was randomly assigned one of the mean LTQ values μ_g of the Baron et al. dataset.
3. Similarly, Sanity estimated the true variances v_g in LTQ to be roughly exponentially distributed in the Baron et al. dataset, with a mean of approximately 2. For the simulated datasets, we assigned each gene a variance in LTQ v_g by drawing a random number from an exponential distribution with mean 2.
4. For each gene g and each cell c, we randomly sampled a log fold-change δ_{gc} from a Gaussian distribution with mean zero and variance v_g.
5. We set the LTQ of gene g in cell c to $\log(\alpha_{gc}) = \mu_{gc} - v_g/2 + \delta_{gc}$.
6. For each cell c we take the total UMI count N_c from one of the cells in the Baron dataset.
7. Finally, for each gene g in each cell c we sample the UMI count n_{gc} from a Poisson distribution with mean $N_c \alpha_{gc}$.

Note that in step 5 we subtracted a term $v_g/2$ from the mean LTQ μ_g before adding the log fold-changes δ_{gc}. This is done to make sure that the expected mean

transcription quotient matches e^{μ_g} and that the sum of the expected transcription quotients is 1, i.e. that we have

$$\sum_g \langle \alpha_{gc} \rangle = \sum_g \langle e^{\mu_g - v_g/2 + \delta_{gc}} \rangle = \sum_g e^{\mu_g} = 1. \tag{4.56}$$

The second simulated dataset was created in exactly the same way, except for the way in which the log fold-changes δ_{gc} were chosen. These were sampled as follows

1. For the first cell $c = 1$, all δ_{g1} are drawn from a Gaussian with mean zero and variance one.

2. For each next cell c a parent cell $\pi(c)$ was assigned, and the log fold-changes δ_{gc} were chosen to be equal to those of its parent cell $\pi(c)$, plus a random Gaussian variable of mean zero and variance one, i.e.

$$\delta_{gc} = \delta_{g\pi(c)} + \theta_{gc}, \tag{4.57}$$

   where θ_{gc} is a Gaussian random variable with mean zero and variance 1.

3. Once every 13 cells, the parent $\pi(c)$ is chosen uniformly randomly from all existing cells, and for all other cells $\pi(c) = c - 1$, i.e. the parent cell is simply the previous cell. Thus, the δ_{gc} are drawn by performing a random walk in gene expression space, where after every 13 random walk steps, a new random walk is started from a randomly chosen cell. This causes the cells to fall on a tree with branches of 13 cells each.

4. Finally, in order to make sure the variance of the LTQs of each gene matches the assigned variance v_g, we calculate the variance $\mathrm{var}(\delta_g)$ of the log fold-changes of gene g along the random walk, and then rescale all the log fold-changes δ_{gc} by multiplying them by $\sqrt{v_g/\mathrm{var}(\delta_g)}$.

Figure 4.9 shows the distributions of the total number of mRNAs per cell, the total number of mRNAs per gene, and the variance in observed mRNA counts for both the Baron dataset and the simulated data. Note that the distributions are highly similarly except for the variances, which are more widely distributed in the simulated data.

4.5.3 Estimating cell-to-cell distances

Given two cells c and c', the squared distance $d^2_{cc'}$ is the sum of the squares of the per gene LTQ differences, i.e.

$$d^2_{cc'} = \sum_g \left(\delta_{gc} - \delta_{gc'}\right)^2, \tag{4.58}$$

where $\mu_g + \delta_{gc}$ is the true LTQ of gene g in cell c and δ_{gc} the true log fold-change. Notably, the true log fold-changes δ_{gc} will differ from the estimated log fold-changes δ^*_{gc} that Sanity reports. In particular, for each gene g and cell c, Sanity also reports an error bar ϵ_{gc} on this estimate.

In order to estimate the distance $d_{cc'}$ while incorporating these error bars we are going to make a few simplifications to make the calculation tractable. First, Sanity estimates an overall variance v_g for each gene and, although this estimate of course also has finite accuracy, we will ignore this uncertainty in v_g and presume that, given the data D_g for gene g, the variance v_g is determined. Second, although there are nonzero covariances across cells in the full posterior distribution of the log fold-changes δ_{gc}, we will ignore these as well and approximate the posterior over all δ_{gc} by a product of independent posteriors with means δ^*_{gc} and variances ϵ^2_{gc}. That is, given v_g and the data D_g for gene g, the posterior for the true log fold-change δ_{gc} is given by the Gaussian:

$$P(\delta_{gc}|v_g, D_g) = \frac{1}{\sqrt{2\pi v_g}} \exp\left[-\frac{(\delta_{gc} - \delta^*_{gc})^2}{2\epsilon^2_{gc}}\right]. \tag{4.59}$$

Given these posteriors and given a pair of cells (c, c') we now want to estimate the sum of the square deviations $(\delta_{gc} - \delta_{gc'})^2$ and to this end we introduce some simplifying notation. Let x_g denote the *estimated* difference in the LTQs of gene g between the two cells, i.e.

$$x_g = \delta^*_{gc} - \delta^*_{gc'}, \tag{4.60}$$

and let Δ_g denote the *true* difference in LTQs:

$$\Delta_g = \delta_{gc} - \delta_{gc'}, \tag{4.61}$$

Also, we denote by $\vec{x}$ and $\vec{\Delta}$ the vectors of estimated and true LTQ differences across all genes.

In order to estimate the true squared deviations Δ^2_g we need a prior distribution $P(\vec{\Delta})$ over the vector of true LTQ differences $\vec{\Delta}$ and a likelihood $P(D_{c,c'}|\vec{\Delta})$ of the data for the two cells given the vector of true LTQ differences between the cells. Once the likelihood and prior are given, the expected square distance $\langle d^2 \rangle$ between the two cells is formally given by the following ratio of integrals

$$\langle d^2 \rangle = \sum_g \langle \Delta^2_g \rangle = \frac{\int d\vec{\Delta} \left(\sum_g \Delta^2_g\right) P(D_{c,c'}|\vec{\Delta}) P(\vec{\Delta})}{\int d\vec{\Delta} P(D_{c,c'}|\vec{\Delta}) P(\vec{\Delta})}. \tag{4.62}$$

To set the prior $P(\vec{\Delta})$ we first note that the overall squared-deviation d^2 is the sum of squared deviations Δ^2_g over all genes. From the entire set of cells Sanity

has estimated that, averaged over all pair of cells, the expected squared deviation of gene g is proportional to v_g, i.e.

$$\langle \Delta_g^2 \rangle = 2v_g. \tag{4.63}$$

Taking this into account, we parametrize the expected square deviations $\langle \Delta_g^2 \rangle$ for a single pair of cells (c, c') by a hyperparameter α which specifies that the expected square deviation for gene g is

$$\langle \Delta_g^2 \rangle = \alpha v_g. \tag{4.64}$$

That is, for $\alpha = 0$ the two cells are expected to have identical expression profiles, whereas for $\alpha = 2$ the two cells are expected to be as distant as a random pair cells.

Noting that the maximum entropy distribution consistent with a given expected square deviation Δ_g^2 is Gaussian, we will assume a Gaussian prior distribution over each Δ_g with mean zero and variance αv_g. That is, given hyperparameter α, we have for the prior

$$P(\vec{\Delta}|\vec{v}, \alpha) = \prod_g \left[\frac{1}{\sqrt{2\pi\alpha v_g}} \exp\left(-\frac{\Delta_g^2}{2\alpha v_g} \right) \right], \tag{4.65}$$

where we have explicitly indicated that this prior depends on the vector $\vec{v}$ of estimated variances v_g and the hyperparameter α. Our final prior $P(\vec{\Delta}|\vec{v})$ is given by using a uniform distribution over the unknown hyperparameter α for each pair of cells, i.e. we have

$$P(\vec{\Delta}|\vec{v}) \propto \int d\alpha P(\vec{\Delta}|\vec{v}, \alpha). \tag{4.66}$$

To obtain the likelihood $P(D_{c,c'}|\vec{\Delta})$, we note that the posterior $P(\delta_{gc}|D_g, v_g)$ is proportional to a product of the likelihood and prior used by Sanity, i.e.

$$P(\delta_{gc}|D_g, v_g) \propto P(D_g|\delta_{gc}, v_g) P(\delta_{gc}|v_g), \tag{4.67}$$

where Sanity's prior is itself a Gaussian with mean zero and variance v_g, i.e.

$$P(\delta_{gc}|v_g) = \frac{1}{\sqrt{2\pi v_g}} \exp\left(-\frac{\delta_{gc}^2}{2v_g} \right). \tag{4.68}$$

Using this, and solving for the likelihood, we find that the likelihood $P(D_g|\delta_{gc}, v_g)$ is given by

$$P(D_g|\delta_{gc}, v_g) = \frac{1}{\sqrt{2\pi}\eta_{gc}} \exp\left(-\frac{(\delta_{gc} - \tilde{\delta}_{gc})^2}{2\eta_{gc}^2} \right), \tag{4.69}$$

with the rescaled mean

$$\tilde{\delta}_{gc} = \delta_{gc}^* \frac{v_g}{v_g - \epsilon_{gc}^2}, \tag{4.70}$$

and variance

$$\eta_{gc}^2 = \epsilon_{gc}^2 \frac{v_g}{v_g - \epsilon_{gc}^2}. \tag{4.71}$$

Finally, combining the likelihoods for δ_{gc} and $\delta_{gc'}$, and defining $x_g = \tilde{\delta}_{gc} - \tilde{\delta}_{gc'}$ and $\eta_g^2 = \eta_{gc}^2 + \eta_{gc'}^2$, we find for the likelihood of the difference in LTQ

$$P(D_{c,c'}|\vec{\Delta}) = \prod_g \left[\frac{1}{\sqrt{2\pi}\eta_g} \exp\left(-\frac{(\Delta_g - x_g)^2}{2\eta_g^2} \right) \right] \tag{4.72}$$

Combining the expressions for the likelihood and prior into equation (4.62), we find the expected square distance can be rewritten as

$$\langle \sum_g \Delta_g^2 \rangle = \frac{\int d\alpha d\vec{\Delta} \left(\sum_g \Delta_g^2 \right) P(D_{c,c'}, \vec{\Delta}|\alpha, \vec{v})}{\int d\alpha d\vec{\Delta} P(D_{c,c'}, \vec{\Delta}|\alpha, \vec{v})}, \tag{4.73}$$

with

$$P(D_{c,c'}, \vec{\Delta}|\alpha, \vec{v}) = \prod_g \frac{1}{2\pi\eta_g\sqrt{\alpha v_g}} \exp\left[-\sum_g \frac{(\Delta_g - x_g)^2}{2\eta_g^2} + \frac{\Delta_g^2}{2\alpha v_g} \right]. \tag{4.74}$$

The integral over $\vec{\Delta}$ in the denominator can be performed analytically to obtain

$$P(D_{c,c'}|\alpha, \vec{v}) = \prod_g \frac{1}{\sqrt{2\pi(\alpha v_g + \eta_g^2)}} \exp\left[-\frac{x_g^2}{2(\alpha v_g + \eta_g^2)} \right]. \tag{4.75}$$

We can use this to define a posterior distribution over α:

$$P(\alpha|D_{c,c'}, \vec{v}) = \frac{P(D_{c,c'}|\alpha, \vec{v})}{\int d\alpha P(D_{c,c'}|\alpha, \vec{v})} \tag{4.76}$$

and rewrite the expectation value for the squared distance as

$$\langle \sum_g \Delta_g^2 \rangle = \sum_g \int d\alpha P(\alpha|D_{c,c'}, \vec{v}) \int d\Delta_g \Delta_g^2 P(\Delta_g|D_{c,c'}, v_g, \alpha), \tag{4.77}$$

with the posterior distribution over Δ_g given α and the data given by

$$P(\Delta_g|D_{c,c'}, v_g, \alpha) = \frac{1}{\eta_g\sqrt{2\pi f_g(\alpha)}} \exp\left[-\frac{(\Delta_g - f_g(\alpha)x_g)^2}{2\eta_g^2 f_g(\alpha)} \right], \tag{4.78}$$

with

$$f_g(\alpha) = \frac{\alpha v_g}{\alpha v_g + \eta_g^2}. \tag{4.79}$$

Finally, performing the integral over all Δ_g we obtain

$$\langle d^2 \rangle = \sum_g \int d\alpha P(\alpha|D_{c,c'}, \vec{v}) \left[x_g^2 f_g(\alpha)^2 + \eta_g^2 f_g(\alpha) \right] . \tag{4.80}$$

Note that, at a given value of α, the expected square difference in the LTQs for gene g is αv_g. If the variance η_g^2 corresponding to the measurement noise in the estimated squared difference x_g^2 is small compared to αv_g, then $f_g(\alpha)$ is approximately 1, and the term in the square brackets is $x_g^2 + \eta_g^2$. That is, in that limit the contribution of gene g to the squared distance is dominated by the estimated change in LTQs x_g^2. In contrast, if the measurement noise is high and $\eta_g^2 \gg \alpha v_g$, the terms in brackets becomes αv_g. That is, in that limit the contribution of gene g is simply the expected squared difference αv_g. The estimated distance is thus given by a weighted average of the terms in the squared brackets over all values of α, weighing each α with its probability $P(\alpha|D_{c,c'}, \vec{v})$ given the data for all genes. Note that, to obtain $\langle d^2 \rangle$ we simply need the weighted averages $\langle f_g \rangle$ and $\langle f_g^2 \rangle$, which we calculate numerically.

4.5.4 Clustering analysis

Assessing the performance of the different normalization methods on downstream clustering of cells is challenging for several reasons. First, for real data the optimal clustering is at best partially known. That is, although we have a number of data sets for which reference clusterings were provided, it is by no means obvious that these reference clusterings are really optimal. Second, clustering is a complex problem for which there is no algorithm that is optimal in all situations, so that the performance may vary depending on what clustering algorithm is used [Kiselev et al., 2019]. Finally, partitions of a set into subsets are fairly complex objects themselves, and it is not even obvious how to quantify the similarity between two possible partitions. Consequently, a host of different similarity measures are used in the literature.

We therefore decided to benchmark the performance of the normalization methods on the clustering by using 6 different scRNA-seq datasets that have annotated reference clusterings, run 3 different popular clustering methods on each, and then use 4 different similarity measures to compare the resulting clusterings with the reference clusterings. We reasoned that the quality of each normalization can be assessed by the robustness in performance across datasets, clustering methods, and similarity measures.

Clustering algorithms

The three clustering algorithms we used were:

- K-means : A method that starts from a random initial clustering and iteratively reassigns objects to clusters so as to minimize intra-cluster variance [Lloyd, 1982].
- Ward: A hierarchical clustering methods that constructs a linkage tree of the objects by iteratively fusing the pair of objects so as to minimize the increase in intra-cluster variance [Ward, 1963].
- Louvain: A more recent method for extracting the community structure of graphs. Starting from a k-nearest neighbor graph, cells are merged in clusters by optimizing the modularity, an objective function that measures the density of links inside communities as compared to links between communities [Blondel et al., 2008]

For both the K-means and Ward method we set the number of clusters to match those of the reference annotation for each dataset (which is not possible for the Louvain clustering). The only parameter for the Louvain clustering is the number of nearest neighbors, which we set to 30.

Clustering similarity measures

Let the sets $\{A\}$ and $\{B\}$ denote two partitions of a set of cells, with $A_i \in \mathbb{N}_+$ and $B_i \in \mathbb{N}_+$ representing the number of the subset to which cell i is assigned in each of the partitions, with $i = 1, ..., C$, and C the number of cells.

The size distributions and the joint distribution of the two partitions are defined as the frequencies

$$P_A(a) = \frac{|\{A_i = a | 1 \leq i \leq C\}|}{C} \tag{4.81}$$

$$P_B(b) = \frac{|\{B_i = b | 1 \leq i \leq C\}|}{C} \tag{4.82}$$

$$P_{AB}(a, b) = \frac{|\{\{A_i = a\} \cap \{B_i = b\} | 1 \leq i \leq C\}|}{C}, \tag{4.83}$$

where $|\cdot|$ denotes the cardinality of a set, i.e. $P_A(a)$ is the fraction of cells belonging to cluster a in partition $\{A\}$, $P_B(b)$ is the fraction of cells belonging to cluster b in partition $\{B\}$, and $P_{AB}(a, b)$ is the fraction of cells that belong to both clusters a and b.

The entropies of these distributions are then defined as

$$H(A) = -\sum_{a\in\mathbb{N}_+} P_A(a)\log P_A(a) \tag{4.84}$$

$$H(B) = -\sum_{b\in\mathbb{N}_+} P_B(b)\log P_B(b) \tag{4.85}$$

$$H(A,B) = -\sum_{a,b\in\mathbb{N}_+} P_{AB}(a,b)\log P_{AB}(a,b), \tag{4.86}$$

and the *mutual information* is defined as

$$I(A;B) = H(A)+H(B)-H(A,B) \tag{4.87}$$

$$= \sum_{a\in A,b\in B} P_{AB}(a,b)\log\frac{P_{AB}(a,b)}{P_A(a)P_B(b)}, \tag{4.88}$$

representing the amount of information the two partitions contain about each other.

As a first measure of similarity, we use the *Normalized mutual information*, defined as

$$NMI(A;B) = \frac{I(A;B)}{\sqrt{H(A)H(B)}}. \tag{4.89}$$

The other three similarity measures are all based on comparing, the reference partition $\{A\}$ with a given other partition $\{B\}$ by counting the number of pairs (i,j) that are either in the same cluster in both partitions (called true positives TP), in the same cluster in the reference $\{A\}$ but not in partition $\{B\}$ (called false negatives FN), in the same cluster in $\{B\}$ but not in the reference (called false positives FP), or that are in different clusters in both partitions (true negatives TN).

The three other similarity measures define similarity in terms of the counts TP, FN, FP, and TN as follows:

- The *Adjusted rand index* or ARI:

$$ARI(A,B) = 2\frac{TP\cdot TN - FN\cdot FP}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)} \tag{4.90}$$

- The *Fowlkes–Mallows index* or FM:

$$FM(A,B) = \sqrt{\frac{TP}{TP+FN}\frac{TP}{TP+FP}}. \tag{4.91}$$

- The *Jaccard index*:

$$Jaccard(A,B) = \frac{TP}{TP+FN+FP}. \tag{4.92}$$

Ranking of the normalization methods based on their similarity scores

Using the 6 annotated datasets, the 3 clustering algorithms and the 4 similarity measures, we obtain a total of 72 different similarity scores for each method (Suppl. Fig. 4.26). To summarize these results, we first calculated the number of times each method was the best performing method across the 72 combinations (Fig. 4.5d). Second, to measure the robustness of each method, we first calculated for each of the 72 combinations and each method m, the ratio $r_m = s_m/s_*$ of the similarity score s_m that method m had, and the highest similarity score s_* on that combination. We then calculated for each method the distribution of the ratio r_m across the 72 combinations (Fig. 4.5e).

4.5.5 Differential expression analysis

Let e_{gc} denote the log-expression of gene g in cell c, C an ensemble of cells, and $\bar{C}$ its complement, i.e. all other cells in the dataset. The t-statistic t_{gC} quantifies the statistical evidence that the average expression of gene g in the cells of set C differs from the average in all other cells:

$$t_{gC} = \frac{\mu_{gC} - \mu_{g\bar{C}}}{\sqrt{\sigma^2_{gC}/|C| + \sigma^2_{g\bar{C}}/|\bar{C}|}} \tag{4.93}$$

$$\mu_{gC} = \frac{1}{|C|}\sum_{c\in C} e_{gc} \tag{4.94}$$

$$\mu_{g\bar{C}} = \frac{1}{|\bar{C}|}\sum_{c\in \bar{C}} e_{gc} \tag{4.95}$$

$$\sigma^2_{gC} = \frac{1}{|C|}\sum_{c\in C} \left(e_{gc} - \mu_{gC}\right)^2 \tag{4.96}$$

$$\sigma^2_{g\bar{C}} = \frac{1}{|\bar{C}|}\sum_{c\in \bar{C}} \left(e_{gc} - \mu_{g\bar{C}}\right)^2, \tag{4.97}$$

with $|C|$ and $|\bar{C}|$ the number of cells in set C and its complement, respectively.

Given a t-statistic t_{gC}, the p-value under a one-sided t-test for the null hypothesis that the gene has the same average expression in set C as in its complement is

$$P(t_{gC}) = \frac{1}{2}\mathrm{Erfc}\left(\frac{t_{gC}}{\sqrt{2}}\right). \tag{4.98}$$

Sorting all genes by the t-statistic t_{gC}, the list of over-expressed genes at a false discovery rate of f is obtained by picking a cut-off t_c such that average of $P(t_{gC})$ for all genes with $t_{gC} > t_c$ is f.

Finally, the reference sets of differentially expressed genes were constructed using a negative binomial generalized linear regression to obtain posterior probability distributions for the class-specific contributions to each gene's expression (also considering contribution of age and sex and a basal expression per gene) (see [Zeisel et al., 2015], Supplementary Materials, Gene expression enrichment analysis).

4.6 Supplementary figures

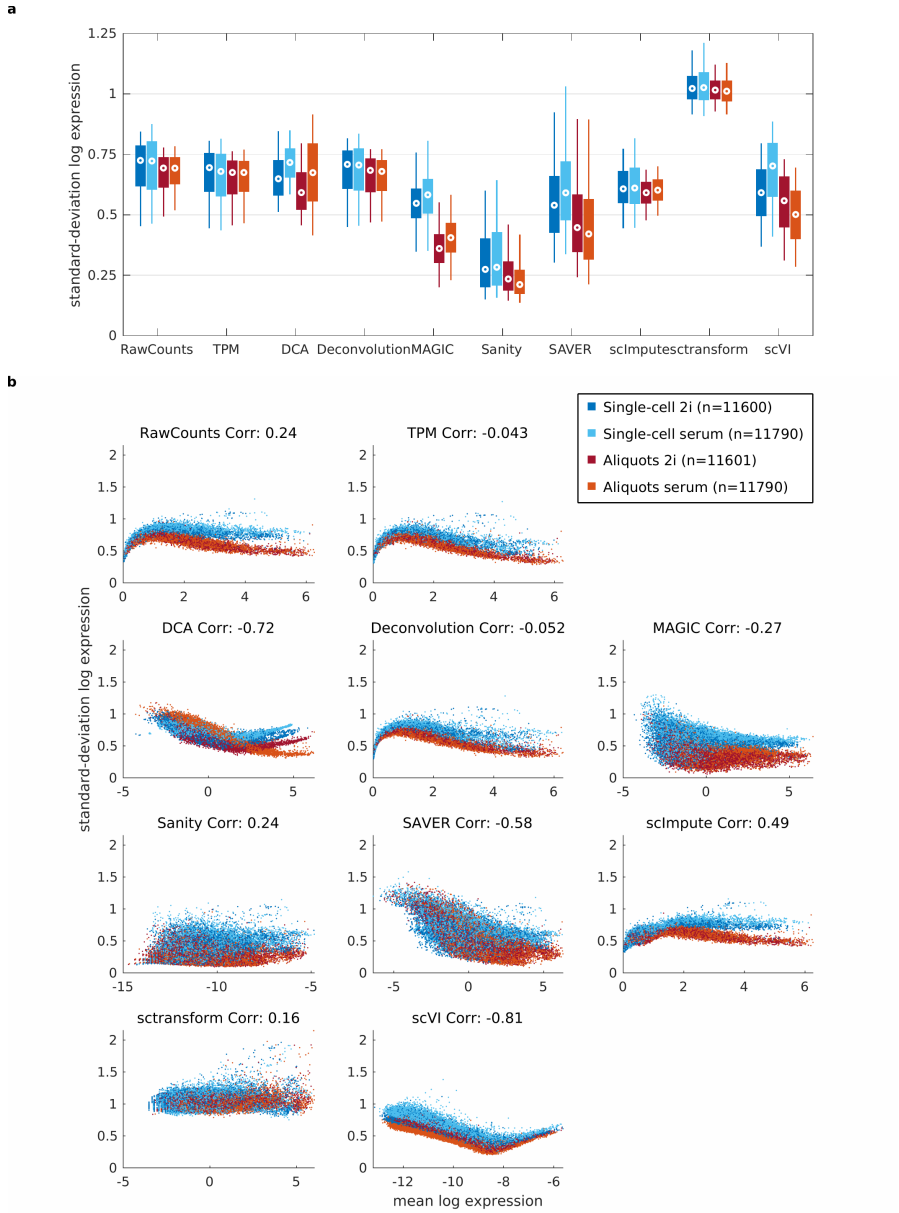


Figure 4.6: **a**: Box-whisker plots showing the median (circle) as well as the 5th, 25th, 75th, and 95th quantiles of the distribution of standard-deviations in log gene expression levels across genes, for each of the 4 datasets (see legend) as inferred by each of the normalization methods. **b** Scatter plots of standard-deviation in log-expression as a function of mean log-expression for all genes in each of the 4 datasets (colors as in panel a) as inferred by each of the normalization methods. The Pearson correlation coefficient between standard-deviation in log-expression and mean log-expression is shown above each plot.

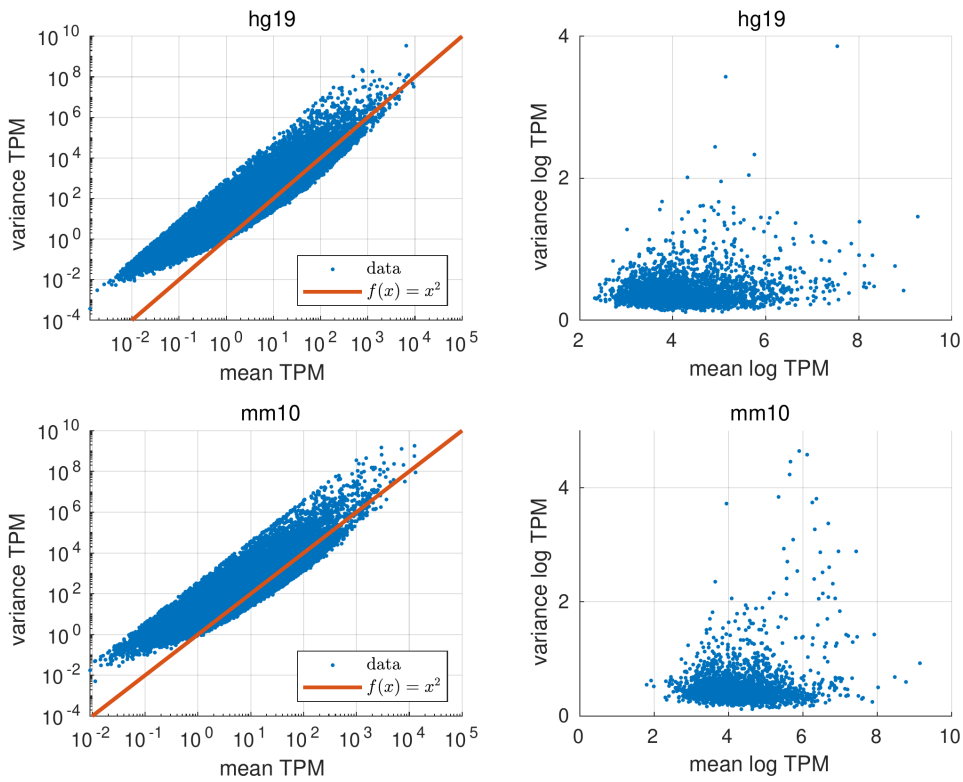


Figure 4.7: In bulk transcriptome data the variance in expression levels scales with the square of the mean, and the variance of log-expression levels is independent of the mean. **Left panels**: Scatter plots of the variance in TPM versus mean TPM across genes for the FANTOM5 expression atlases for human (top) and mouse (bottom) based on deep CAGE sequencing [Forrest et al., 2014]. Each dot corresponds to one promoter. Both axes are shown on logarithmic scales and the red line shows the quadratic relationship $y = x^2$. **Right panels**: The same scatter plots but now showing the variance in log-TPM versus the mean in log-TPM. Note that there is no longer any systematic dependence between mean and variance.

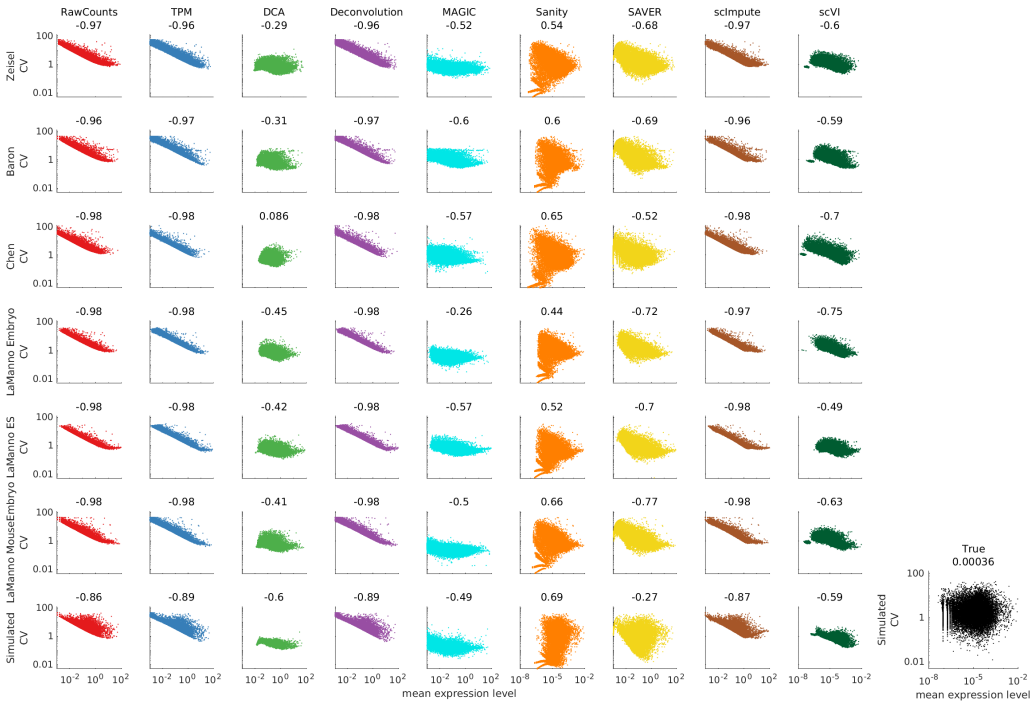


Figure 4.8: Scatter plots of the CV against mean of expression levels across genes. Rows correspond to different scRNA-seq datasets. Colors and columns correspond to the different methods used to normalize the data, as indicated above each column together with the Pearson correlation between log(CV) and log(mean). The bottom row corresponds to the simulated dataset and the black scatter on the right shows the true means and CVs used in the simulation.

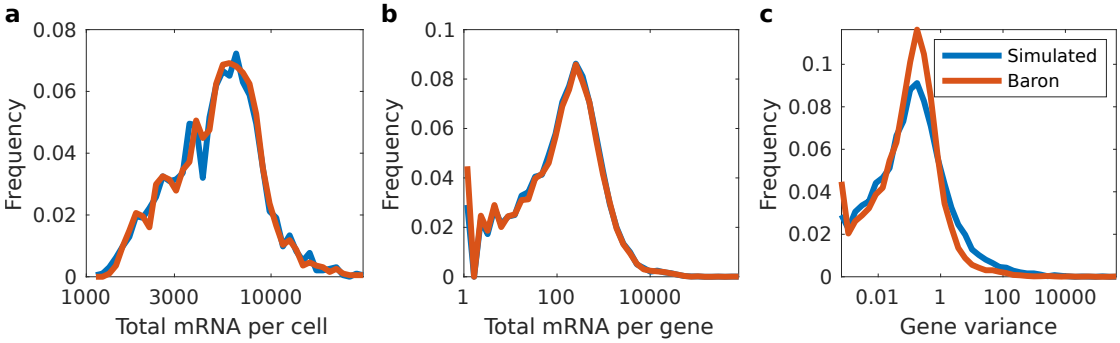


Figure 4.9: Expression statistics of the simulated dataset match those of the *Baron* dataset. **a**: Distribution of total number of UMI per cell N_c in the simulated dataset (blue) and the *Baron* dataset (red). **b**: Distribution of total number of UMI per gene N_g in the simulated dataset (blue) and the *Baron* dataset (red). **c**: Distribution of variance per gene calculated on the raw count matrix obtained from the simulated dataset (blue) and the *Baron* dataset (red).

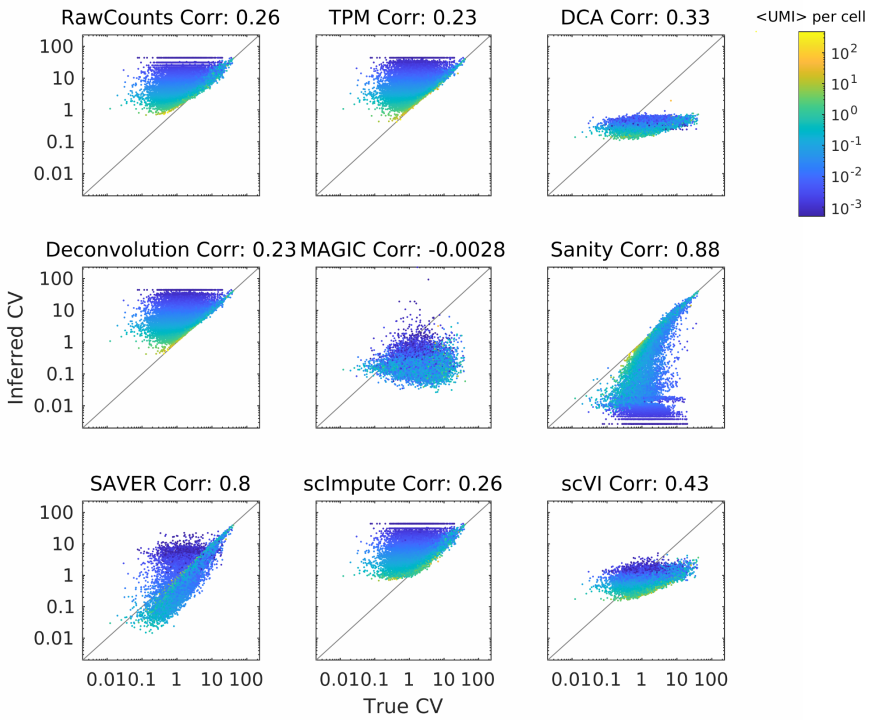


Figure 4.10: Comparison of the true CVs in expression used to simulate the data and the CVs in expression inferred by each of the normalization methods on the simulated dataset. Each panel shows a scatter plot of the true CV (horizontal axis) against the CV inferred by the normalization method (vertical axis) across all genes. The color of each data point shows the mean expression level of the gene (average number of UMI per cell, see color bar). Both axes are shown on a logarithmic scale. The Pearson correlation between the inferred CVs and the true CVs is shown on top of each panel.

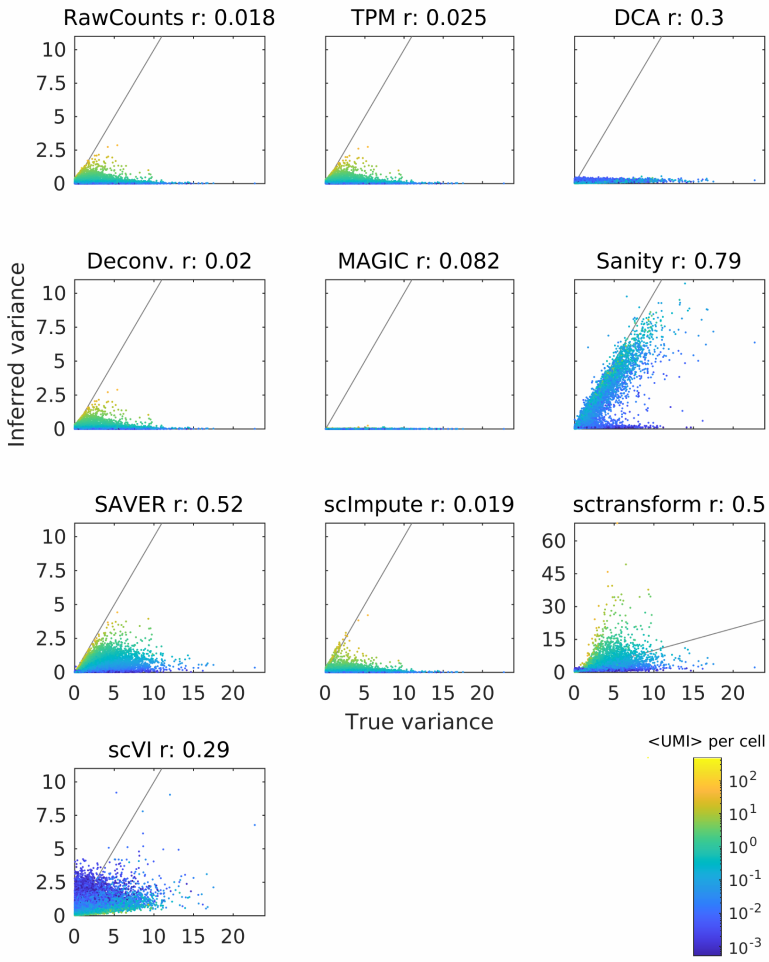


Figure 4.11: Comparison of the true variances in log-expression used to simulate the data and the variances in log-expression inferred by each of the normalization methods on the simulated dataset. Each panel shows a scatter plot of the true variance (horizontal axis) against the variance as inferred by the normalization method (vertical axis) across all genes. The color of each data point shows the mean expression level of the gene (average number of UMI per cell, see color bar). The Pearson correlation r between the inferred variances and the true variances is shown above each panel. Note that for sctransform the inferred variance corresponds to the variance of the z-values.

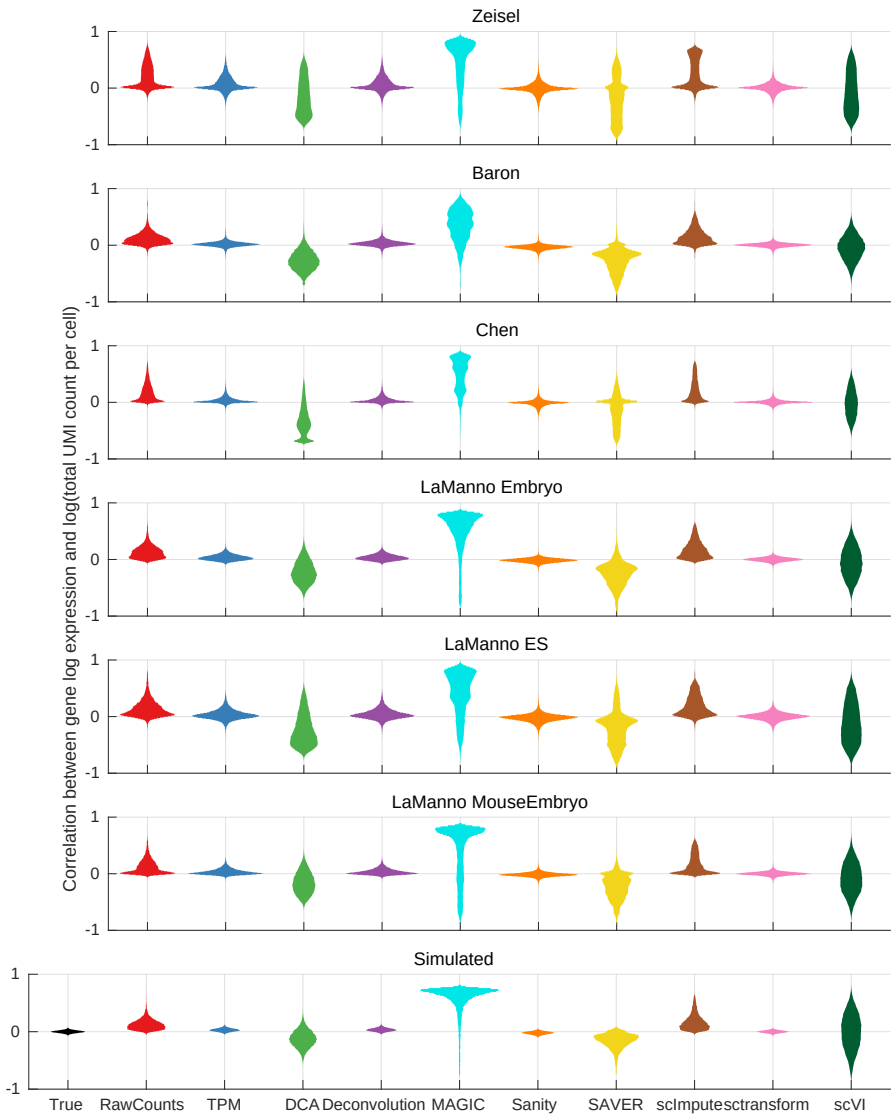


Figure 4.12: Violin plots of the distributions of correlation coefficients between inferred log-expression levels of genes and log of total UMI count per cell. Rows correspond to different datasets, as indicated above each row, with the bottom panel corresponding to the simulated dataset. Columns correspond the different methods, as indicated at the bottom.

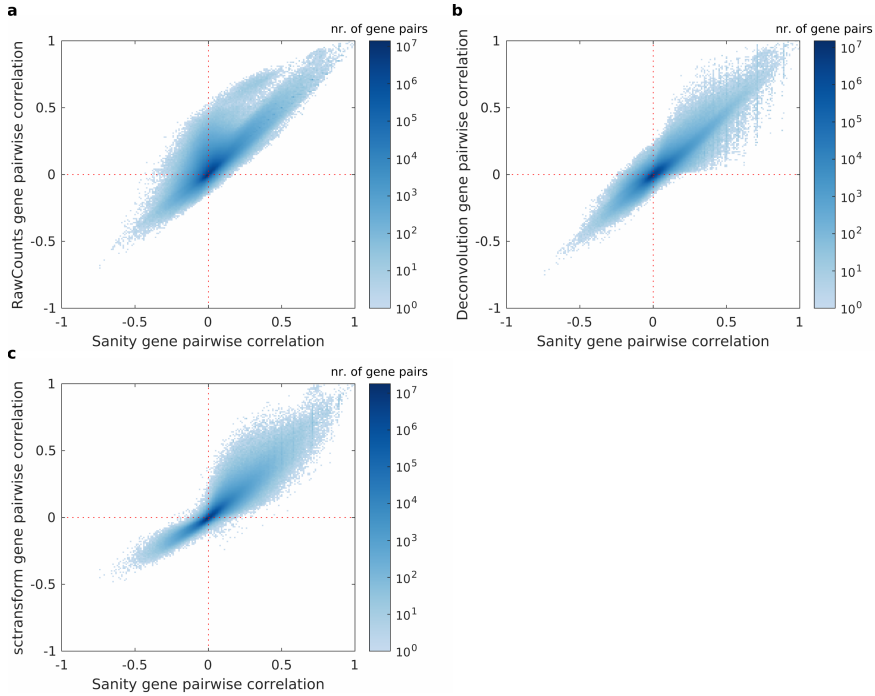


Figure 4.13: Density plots of the Pearson correlations of normalized log-expression values of all pairs of genes as inferred by Sanity (x-axis) and the corresponding pair correlation as inferred by RawCounts (**a**), Deconvolution (**b**), and sctransform (**c**) on the y-axis, for the Baron dataset. The color scale shows the density in $\log_{10}$ of gene pairs and values $\log_{10}(0)$ are shown in white.

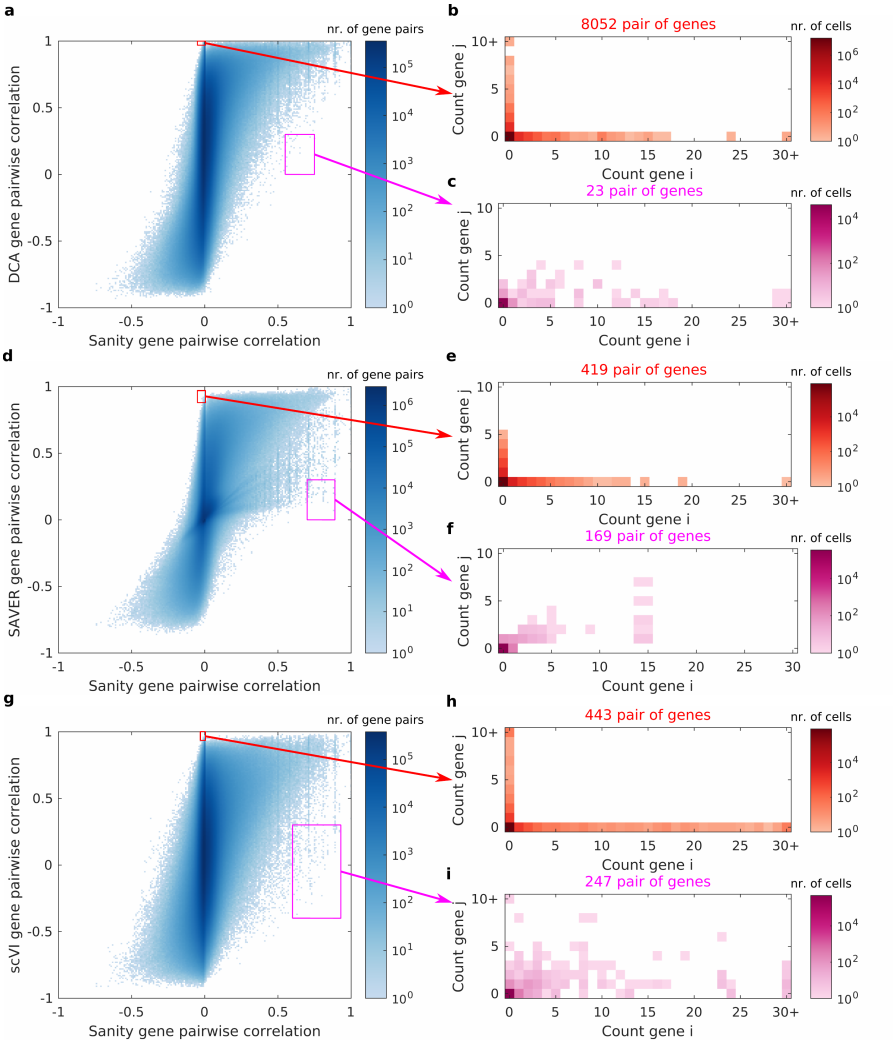


Figure 4.14: Density plots of the Pearson correlations of normalized log-expression values of all pairs of genes as inferred by Sanity (x-axis) and the corresponding pair correlation as inferred by DCA (**a**), SAVER (**d**), and scVI (**g**) on the y-axis, for the Baron dataset. The color scale shows the density in $\log_{10}$ of gene pairs and values $\log_{10}(0)$ are shown in white. For panels **a**, **d**, and **g**, the red and magenta rectangles show selections of gene pairs for which the two methods disagree most strongly on the correlation. For each such set of pairs, we counted the number of times $n_{i,j}$, across all pairs and all cells, for which i UMI were observed for the first gene and j UMI for the second gene. The panels **b**, **c**, **e**, **f**, **h**, and **i** show the corresponding 2-dimensional histograms $n_{i,j}$ for each selected set with the number of pairs indicated above the panel. The height of the histogram is shown in $\log_{10}$ as a color and values $\log_{10}(0)$ are shown in white.

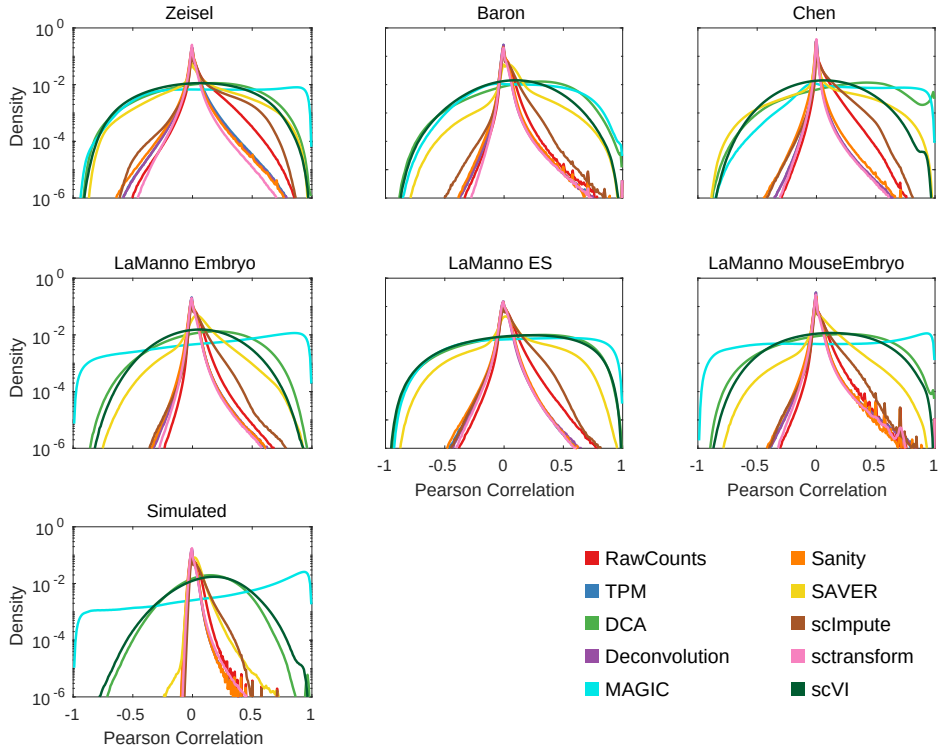


Figure 4.15: Distributions of the Pearson correlations of all pairs of genes, as inferred by each normalization method. Each panel corresponds to one dataset (indicated above it) and each color corresponds to one of the normalization methods, indicated in the legend. Note that the y-axis is shown on a logarithmic scale. Methods that map expression states to a low-dimensional representation have wide distributions of correlations, whereas methods that do not have correlations sharply peaked around zero.

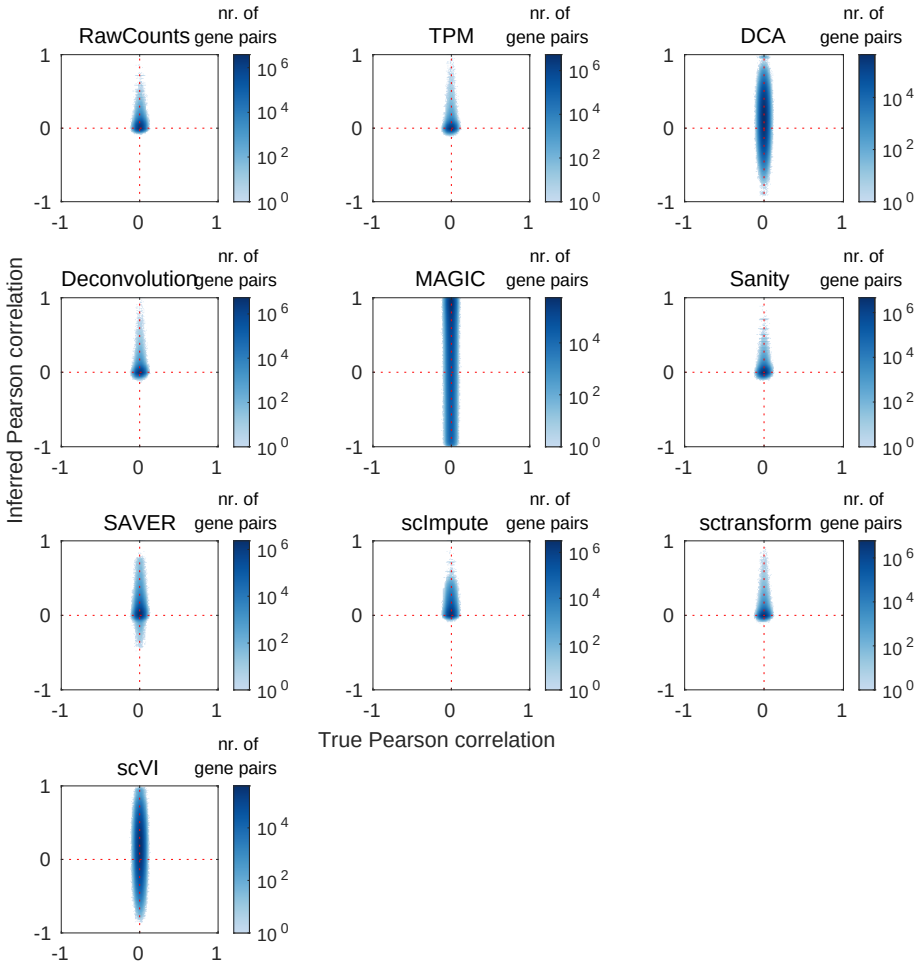


Figure 4.16: Many methods generically predict spurious correlations on simulated data without any true correlations in the expression patterns of genes. The density plots show the Pearson correlations of normalized log-expression values of all pairs of genes as inferred by each method (y-axis) versus the true correlations used in the simulation (which are all near zero). Each panel corresponds to the method indicated at the top. The color scale shows the density in log_{10} of gene pairs and values $\log_{10}(0)$ are shown in white.

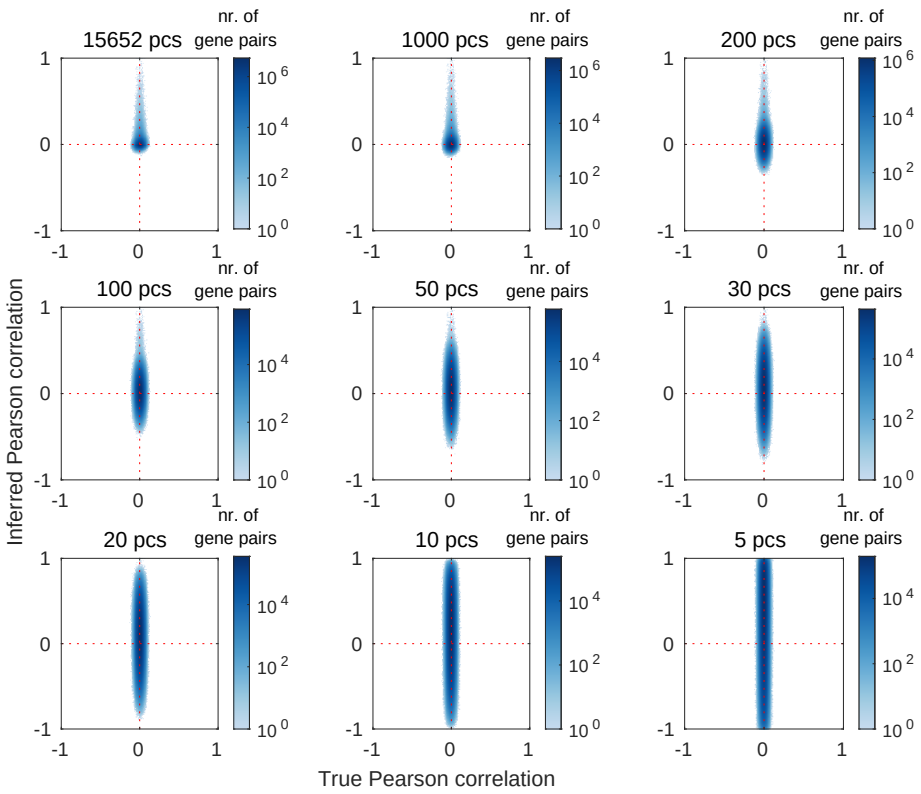


Figure 4.17: Projecting gene expression patterns on the top n PCA components introduces spurious gene expression correlations. The density plots show the Pearson correlations, for all pairs of genes, of the log-expression values that result from projecting the TPM normalized expression values on the first n PCA components (y-axis) versus the true correlations used in the simulation (which are all near zero). Each panel corresponds to a different number n of top PCA components, going from all genes (top right) to only 5 components at the bottom right. The color scale shows the density in $\log_{10}$ of gene pairs and values $\log_{10}(0)$ are shown in white.

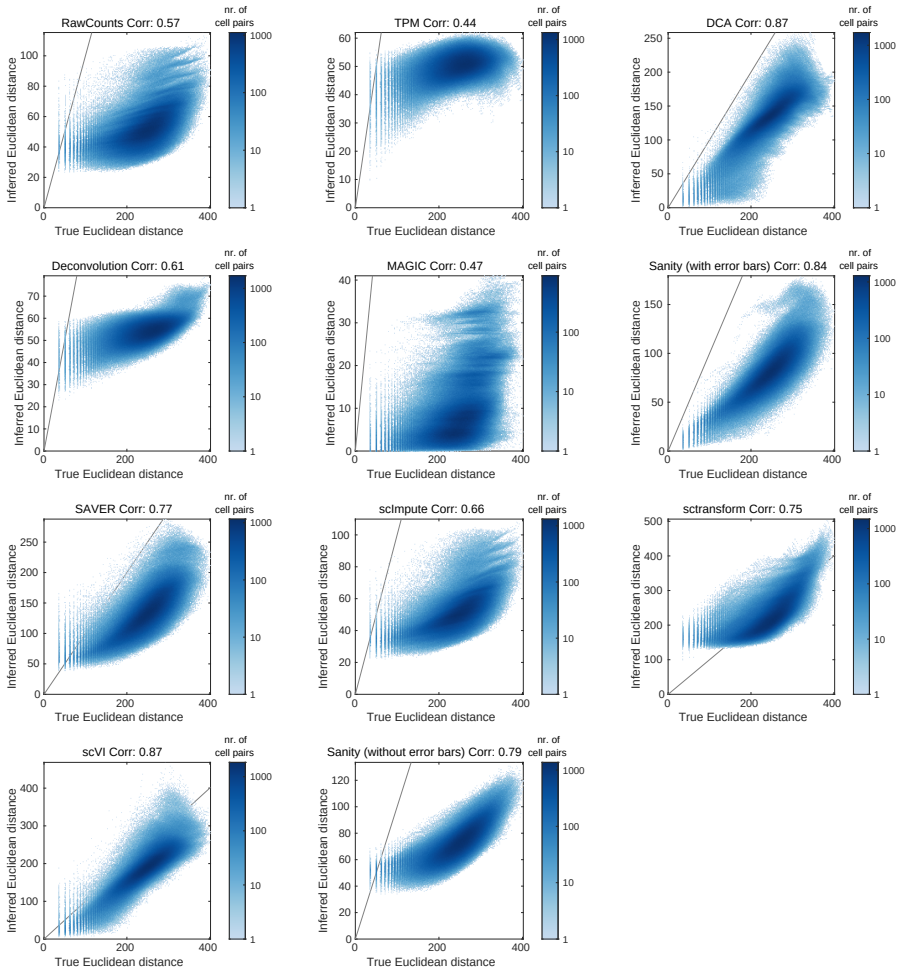


Figure 4.18: Comparison of the true distances between all pairs of cells with the distances estimated by each method, using all genes. The panels show density plots of the Euclidean distances of the true LTQ vectors of each pair of cells (horizontal axis) versus the Euclidean distances of the estimated log-expression vectors. Each panel corresponds to one method. The top of each panel indicates the method and the overall Pearson correlation between the true and estimated distances. For reference, each panel also shows the line $y = x$.

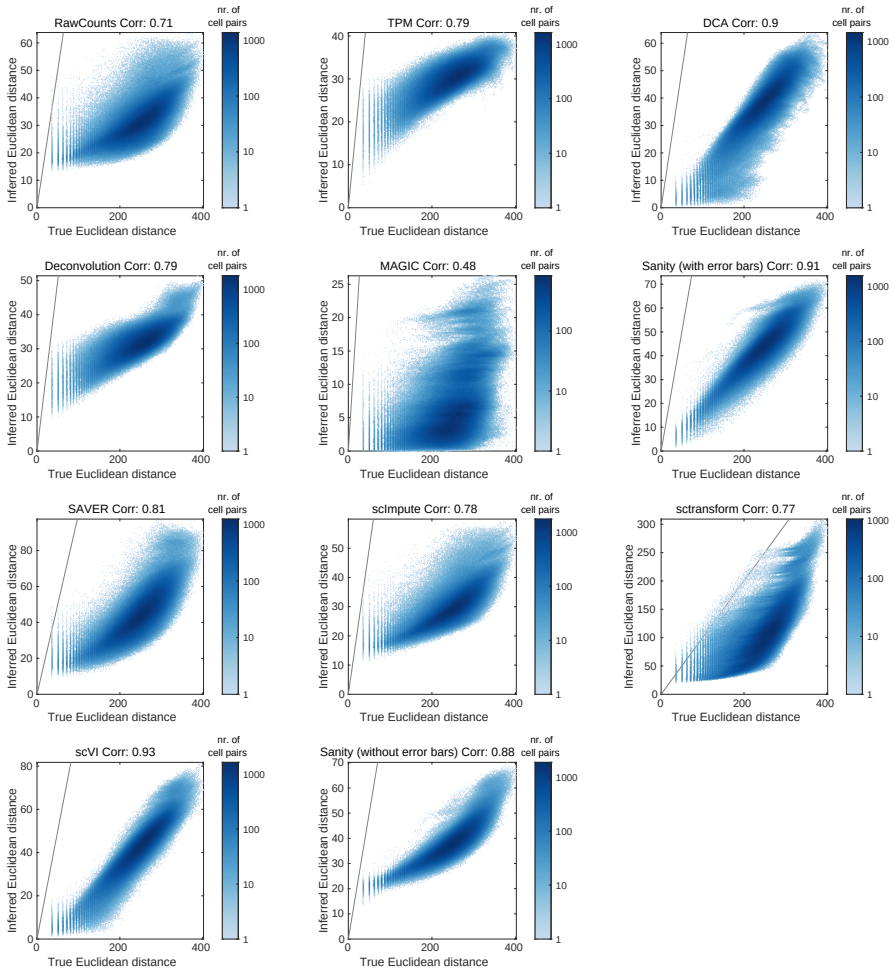


Figure 4.19: Comparison of the true distances between all pairs of cells with the distances estimated by each method, using only genes with at least 1 UMI per cell on average. The panels show density plots of the Euclidean distances of the true LTQ vectors of each pair of cells (horizontal axis) versus the Euclidean distances of the estimated log-expression vectors. Each panel corresponds to one method. The top of each panel indicates the method and the overall Pearson correlation between the true and estimated distances. For reference, each panel also shows the line $y = x$.

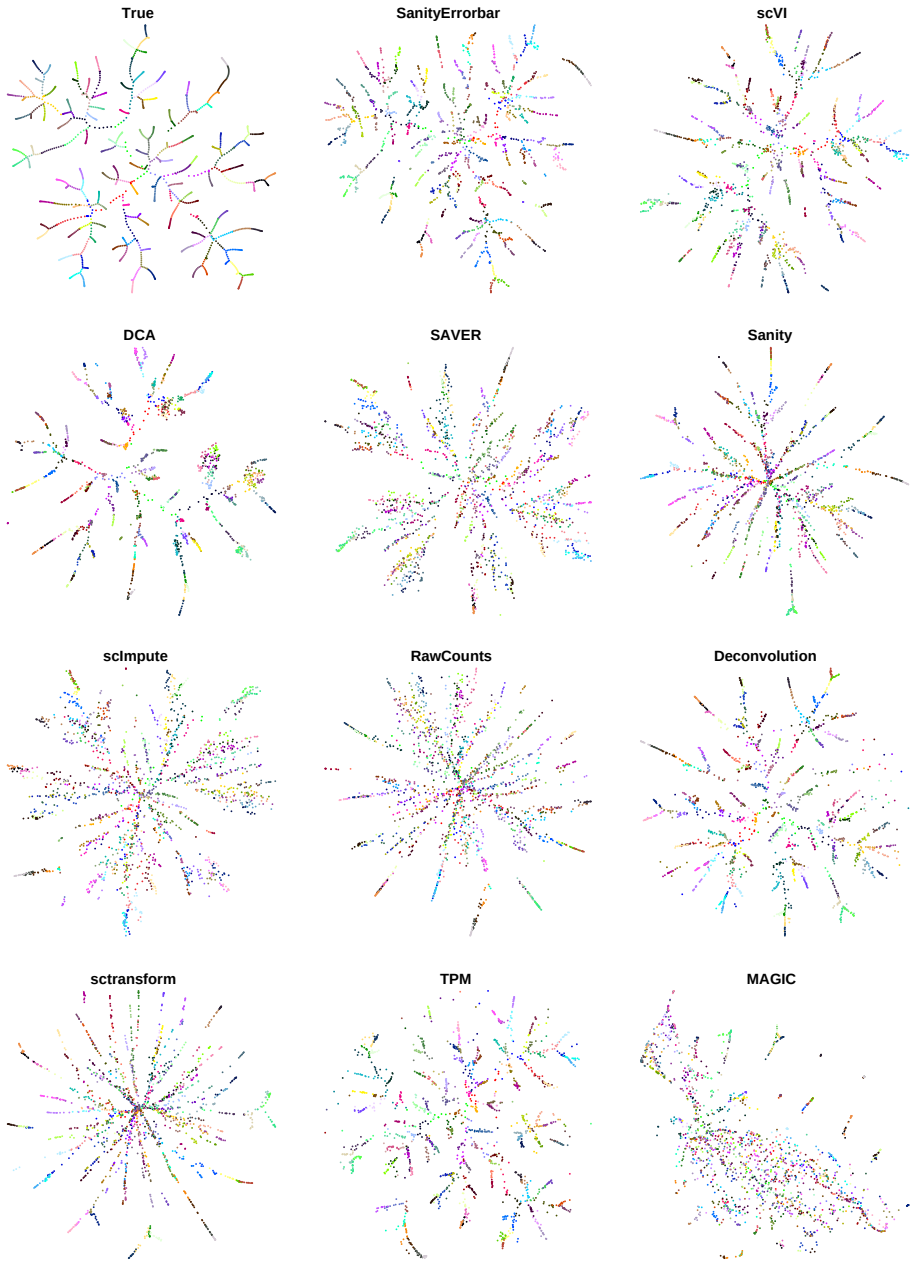


Figure 4.20: T-SNE visualizations of the true and estimated distances between pairs of cells for the simulated dataset corresponding to a branching random walk in gene expression space. The top left panel shows a t-SNE visualization of the true distances between the cells, with each 13-cell branch of the random walk given a different color as a guide for the eye. For each method, t-SNE was run with the same parameters, and starting from an initial condition corresponding to the visualization of the true distances, using the matrix of cell-to-cell distances as estimated from the normalized log-expression values. Methods are sorted from top left to bottom right by the accuracy of their k nearest-neighbor predictions, i.e. the area under the curve of the fraction of correct nearest-neighbors as a function of k.

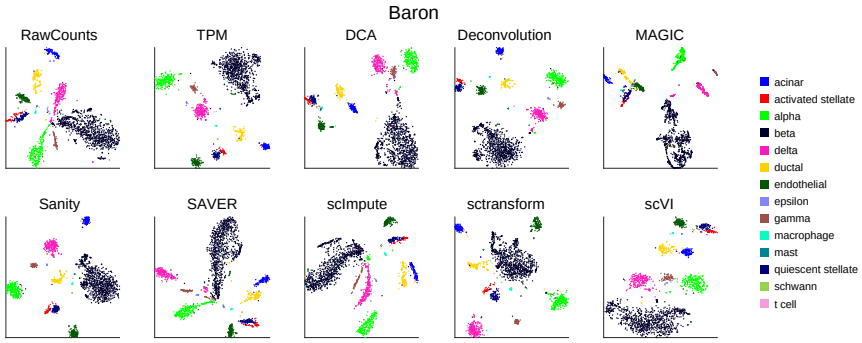


Figure 4.21: T-SNE visualizations of the Baron dataset. Each panel shows a t-SNE visualization of the Baron dataset using the normalized gene expression values of the method indicated at the top. Each point represents a cell and is colored by the cell type annotated in the original publication.

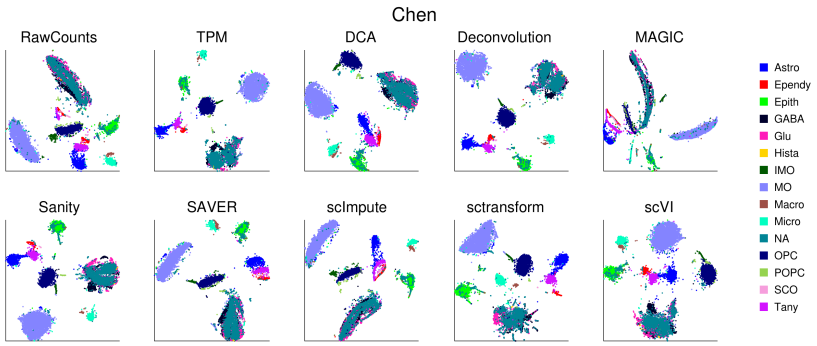


Figure 4.22: T-SNE visualizations of the Chen dataset. Each panel shows a t-SNE visualization of the Chen dataset using the normalized gene expression values of the method indicated at the top. Each point represents a cell and is colored by the cell type annotated in the original publication.

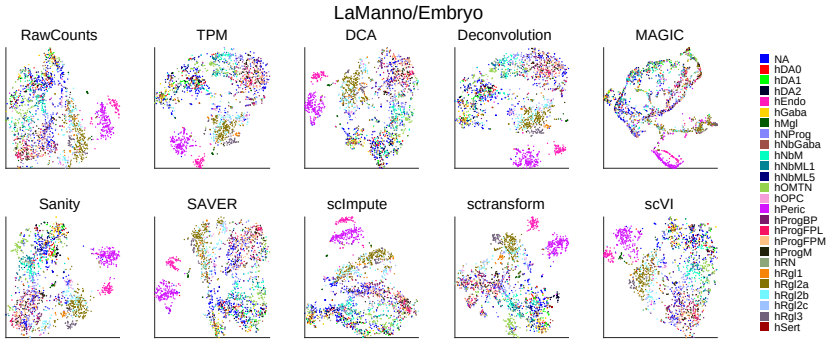


Figure 4.23: T-SNE visualizations of the LaManno/Embryo dataset. Each panel shows a t-SNE visualization of the LaManno/Embryo dataset using the normalized gene expression values of the method indicated at the top. Each point represents a cell and is colored by the cell type annotated in the original publication.

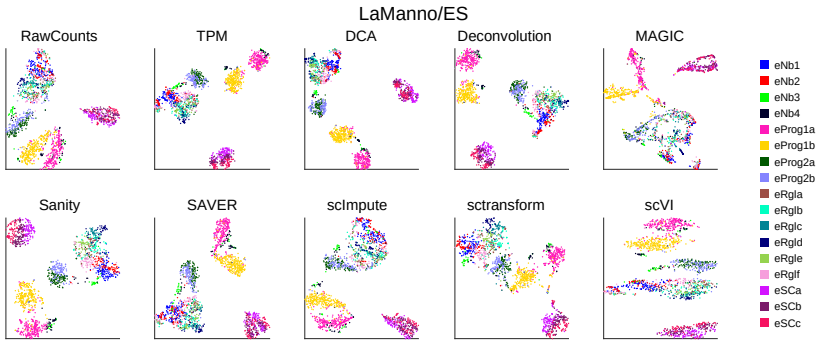


Figure 4.24: T-SNE visualizations of the LaManno/ES dataset. Each panel shows a t-SNE visualization of the LaManno/ES dataset using the normalized gene expression values of the method indicated at the top. Each point represents a cell and is colored by the cell type annotated in the original publication.

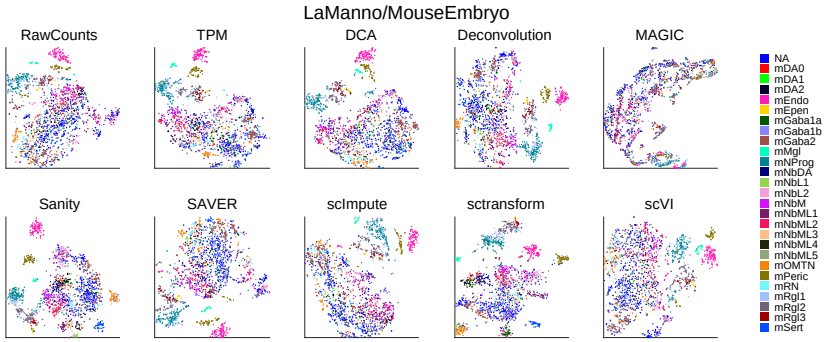


Figure 4.25: T-SNE visualizations of the LaManno/MouseEmbryo dataset. Each panel shows a t-SNE visualization of the LaManno/MouseEmbryo dataset using the normalized gene expression values of the method indicated at the top. Each point represents a cell and is colored by the cell type annotated in the original publication.

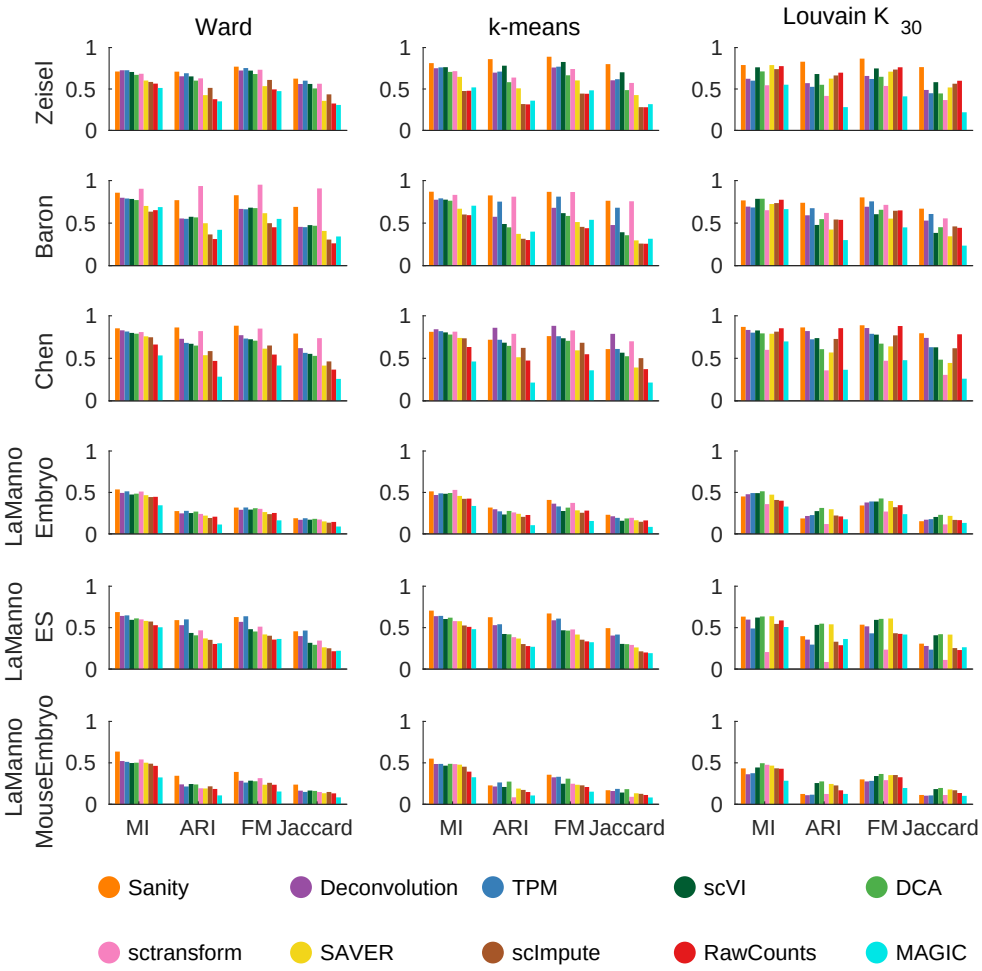


Figure 4.26: Similarities between the reference clusters and the clusters inferred using the normalized gene expression values of the different methods across 6 datasets (rows), 3 clustering methods (main columns), and 4 different similarity metrics (columns within each main column). Clustering was carried out using either hierarchical clustering with Ward's method (left column) k-means clustering (middle column) or Louvain clustering with 30 nearest-neighbors (right column). The similarity measures used were the mutual information (MI), the Adjusted Rand Index (ARI), the Fowlkes–Mallows index (FM), and the Jaccard index. Each similarity measure takes values between 0 (no similarity) and 1 (perfect match). Each group of bars shows the similarity scores for a particular combination with colors indicating the different methods (see legend). Methods are sorted from left to right according to their average similarity score across all combinations.

4.7 Supplementary Text 1: Additional properties of Sanity's model

4.7.1 Sanity outperforms other methods in identifying differentially expressed genes

As another example of downstream analysis we consider the ability of the normalized expression values to identify genes that are upregulated in particular subtypes of cells. That is, we aim to identify genes whose average expression in a given subtype of cells is significantly higher than its average in all other cells. A simple and standard statistic for comparing the averages of populations is the t-statistic and we used this to identify upregulated genes for each cell type in a given dataset. In particular, for each gene g and each cell type k annotated in a given dataset, we calculated a t-statistic

$$t_{gk} = \frac{\mu_{gk} - \mu_{g\bar{k}}}{\sqrt{\sigma^2_{gk}/n_k + \sigma^2_{g\bar{k}}/n_{\bar{k}}}}, \tag{4.99}$$

where μ_{gk} is the average of the normalized expression values of gene g in cells of type k, $\mu_{g\bar{k}}$ is the average in all other cells, σ^2_{gk} and $\sigma^2_{g\bar{k}}$ the corresponding variances in normalized expression levels, and n_k and $n_{\bar{k}}$ the number of cells in type k and the number of all other cells. The t-statistic t_{gk} quantifies the statistical evidence that gene g's average expression in cell type k is higher than in the other cells. To predict a set of upregulated genes, one would then pick a cut-off in t-statistic corresponding to a particular rate of false discovery (FDR), e.g. a 5% FDR. By applying this procedure to the normalized expression values of each method we derived, for each method, a set of upregulated genes for each cell type k of a given dataset of interest.

To test the performance of these predicted sets of upregulated genes we compared these lists with similar lists of predicted upregulated genes from the original publications. For 3 of our test datasets, i.e. the Zeisel and two LaManno datasets, the authors published, for each identified cell type, a list of genes that had higher average expression in the cell type compared to the other types of cells [Zeisel et al., 2015, La Manno et al., 2016]. These lists were obtained using a fairly complex regression procedure and it is of course debatable whether these published lists can be treated as a gold standard. However, since they were obtained using a method that is very different from our simple t-statistic, we reasoned that the match to these reference lists can still be used to assess the relative performance of the different normalization methods.

For each normalization method we calculated a precision-recall curve by producing one sorted list of the t-statistics t_{gc} for all genes in all subtypes and, as a function of a cut-off on t, compared the predicted set of significantly upregulated genes, with the reference lists published in the original study. Figure 4.27 shows the precision-recall curves obtained for each of the methods on each of the

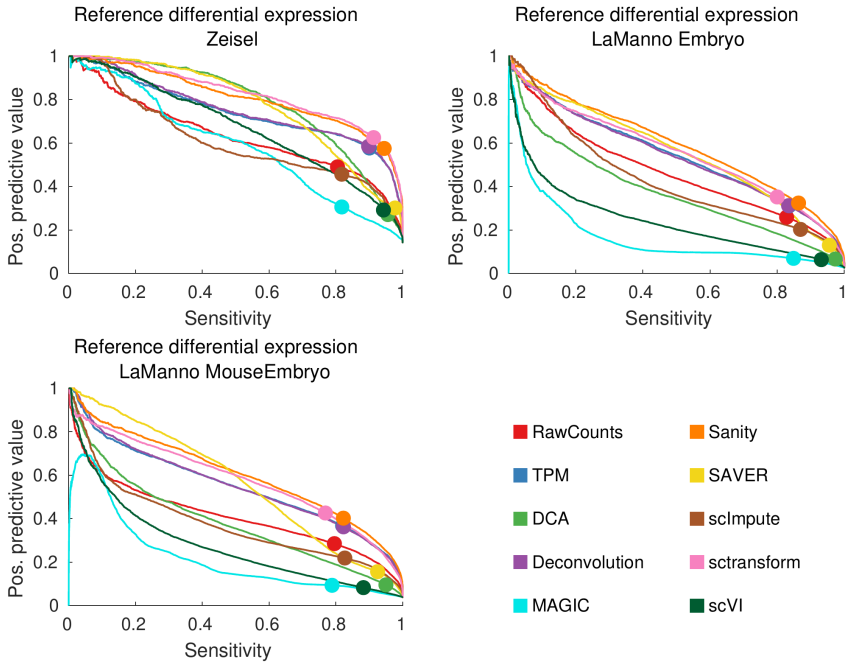


Figure 4.27: Precision recall curves showing the positive predictive value, i.e. the fraction of predicted upregulated genes that correspond to upregulated genes in the corresponding reference list, as a function of sensitivity, i.e. fraction of all genes in the reference lists that were predicted, as obtained using the t-statistics for each of the normalization methods (colors, see legend) for the Zeisel (top left panel) and two LaManno datasets (top right and bottom left panels). The dots show the values that are obtained when using a cut-off on the t-statistic corresponding to a false discovery rate of 5% (based on a one-sided t-test, see Materials and Methods).

3 datasets for which reference lists were available. The colored dots indicate the sensitivity and positive predictive values (PPV) that are obtained for each method when using a t-statistic cut-off corresponding to a 5% FDR. We see that, for each dataset, Sanity achieves the highest accuracy of predictions, i.e. a higher PPV at a given sensitivity than all other methods, followed by sctransform, and then the simple TPM and Deconvolution methods. Note that, at a 5% FDR, the more complex DCA, MAGIC, and scVI methods all predict very large numbers of up-regulated genes which leads to low PPVs. In summary, these results suggest that Sanity's normalized expression levels also achieve highest accuracy for downstream identification of differentially regulated genes.

4.7.2 Limitations of Sanity's model and going beyond them

To motivate Sanity's model we started from the simplest model of gene expression, in which there is a constant rate of transcription λ, and constant rate of mRNA decay μ, leading to a Poisson distribution of the number of mRNAs with average $a = \lambda/\mu$. One popular way in which this simple model can be generalized is to assume that the transcription rate λ can stochastically switch between an 'on' state with high transcription rate λ_{on} and an 'off' state with low/basal transcription rate λ_{off}. However, in reality the transcription rate at a given promoter can almost certainly take on many more than two values. The transcription rate is likely a complex function of the chromatin state and binding configurations of transcription factors at both the promoter and enhancers, and of the 3D structure of the chromosome. In addition, transcription and decay rates will also vary in a continuous manner as a function of the concentrations of RNA polymerases, basal transcription factors, RNAses, and even of the cell cycle state, size of the nucleus, and so on. Recognizing this, we thus generalized the stochastic model of gene expression to assume that, for each gene g in each cell c, there are unknown time-dependent functions $\lambda_g(t)$ and $\mu_g(t)$ that set the transcription rate and mRNA decay rate for gene g at a time t in the past of the time at which the cell was sampled for measurement. Under that model, the expected number of mRNAs $\langle m_{gc} \rangle$ for gene g in cell c is given by the integral

$$\langle m_{gc} \rangle = \int_0^\infty \lambda_{gc}(t) \exp\left[-\int_0^t \mu_{gc}(s) ds \right] dt \equiv a_{gc}, \tag{4.100}$$

which we defined as the transcription activity a_{gc}. The actual number of mRNAs m_{gc} is then a Poisson sample with expected value a_{gc}.

Note that one should think of the functions $\lambda_g(t)$ and $\mu_g(t)$ as fluctuating on a typical time scale. That is, in the model both transcription and mRNA decay events are effectively described as instantaneous events, i.e. $\lambda_g(t)dt$ is the probability that a transcript is produced during the small time interval between t and $t + dt$ in the past. However, the actual process of transcribing a gene and processing the pre-mRNA into an mRNA takes a considerable amount of time, especially for long genes. One can thus think of $\lambda_g(t)dt$ as the probability that an mRNA is finished during the time interval length dt. This probability will be a convolution of probability distribution of the time τ that it takes to transcribe and process, and the transcription *initiation* rate at time $t - \tau$.

Similar remarks apply to the fluctuating mRNA decay rate $\mu_g(t)$. Note also that, because different mRNAs of gene g may be in different states, e.g. bound by different RNA binding complexes, the mRNA decay rate of different mRNAs of the same gene may be different and one should think of $\mu_g(t)$ as the average decay rate of all mRNAs of gene g at time t. In summary, our general model will apply when during short time intervals dt that are long relative to the time it takes for

a transcription initiation event to occur, but short relative to the time between transcription initiation events and the life times of mRNAs, there is effectively a constant probability for a transcription event or mRNA decay event to occur. Under those assumptions the statistics of the mRNA numbers at any time point are still a Poisson distribution with mean the transcription activity a_{gc} defined above.

Although we believe this model is consistent with current biophysical knowledge about the gene expression process, it is certainly possible to imagine biophysical processes that could not be described by this model. To give one example, it is conceivably that at some point in the near future it will be discovered that large complexes exist that contain collections of RNA polymerases together with general transcription factors such that, when such a complex binds to a promoter, a process is started by which all the RNA polymerases in the complex are fed in a regular manner onto the promoter so as to each start a new transcript at regular time intervals until all RNA polymerases in the complex are used up. Clearly, during that period of time we cannot describe the process as having only a certain probability per unit time for a transcript to be produced. One might still argue that one could simply 'zoom out' to a larger time scale at which the processing of the entire complex appears as a single event in which a burst of transcripts is initiated. But the size distribution of these bursts will then be given by the size distribution of these complexes containing RNA polymerases, and these may not respect Poisson statistics. As another example, the model also effectively assumes that the probability of one mRNA to decay during a short time interval dt is proportional to the total number of mRNAs for gene g. If, for example, mRNAs for gene g were to reliably aggregate together into complexes, and all mRNAs within such complexes were degraded at the same time, then this assumption could also break down. We give these examples not because we believe they likely apply, but just to stress that the biophysical processes involved in gene expression are so complex that we cannot exclude that processes exist that strongly violate our model's assumptions.

Another assumption that our model makes is that fluctuations in the LTQ of one gene do not significantly effect the LTQs of other genes. As discussed in the supplementary methods, this assumption is likely accurate as long as not one or a few genes are responsible for a large fraction of all mRNAs in the cell. However, we have also noted that, although rare, in practice there can be cases where a single gene is responsible for a large fraction of the reads in one cell, e.g. hemoglobin can be responsible for more than 50% of the reads in red blood cells.

Information from intronic sequencing reads

A notable limitation of our model is that it effectively ignores the process of transcription itself, including splicing and other mRNA maturation processes. In par-

ticular, the model currently only considers the number of unique mRNA molecules n_{gc} that were sequenced for each gene g in each cell c, and ignores sequencing reads coming from introns. However, the amount of intronic reads may give important information about the *time derivative* of the gene expression state [La Manno et al., 2018, Bergen et al., 2020] which can be especially useful when aiming to reconstruct the trajectories that cells follow through gene expression space. Thus, an important future extension of Sanity's model is to include information from intronic reads.

Incorporating information about the cell's total transcription activity

In our current model we focus on the relative transcription activities, i.e. the transcription quotients α_{gc}, and treat the product of the total transcription activity of a cell A_c and the capture and sequencing efficiency p_c as one effective parameter λ_c that is integrated out of the likelihood. Depending on the system under study and the particular scRNA-seq protocol used, the variation in the capture and sequencing efficiency p_c across cells may be larger than the fluctuations in total activity A_c, and in those cases the variation in total UMI counts N_c will mainly reflect technical variability and not contain much useful biological information. Moreover, some of the variations in A_c may correspond to stochastic fluctuations in cell size that have no effect on cell state. However, especially in more complex tissues there may be cells of different types that have very different sizes and total activities A_c. In those cases the total UMI count N_c may well contain useful information about the cell type of cell c and one might want to make use of this information. Note, however, that it is easy to simply use the total counts N_c as a separate quantity in addition to the estimated LTQs $\mu_g + \delta_{gc}$.

4.7.3 The relation to negative binomial noise models and zero-inflation

Virtually all statistical models of scRNA-seq data agree that the observed UMI counts are a Poisson sample of an underlying expected UMI count $\langle n \rangle$. If one assumes that this expected count is completely constant across cells, then the fraction of cells with zero counts should match Poisson statistics, i.e. equal $e^{-\langle n \rangle}$. In reality much higher fractions of zeroes are observed and this phenomenon is known as zero-inflation. However, there is of course no reason to assume that the expected UMI count $\langle n \rangle$ must be constant across all cells, i.e. this expectation will vary from cell to cell due to changes in gene expression state, cell size, mRNA capture efficiency, and sequencing depth. Once such variations are taken into account, evidence of zero-inflation disappears [Svensson, 2020]. Since our model explicitly takes such variations into account, there is thus no need to incorporate any zero-inflation.

Since all models agree that the observed UMI count n is a Poisson sample of the expected UMI count $\langle n \rangle$, the distribution of UMI counts is a convolution of

the Poisson distribution with the distribution of expected UMI counts $\langle n \rangle$ across cells. The negative binomial model arises if one assumes that $\langle n \rangle$ follows a gamma distribution across cells. However, as far as we are aware, there is no reason to assume that the expected counts $\langle n \rangle$, which depend on variations in transcription quotient, cell size, capture efficiency, and sequencing depth, should follow a gamma distribution. The gamma distribution is assumed mainly for mathematical convenience, since the convolution with the Poisson can then be calculated analytically, yielding a negative binomial distribution.

In contrast, we have argued that the quantities that we want to estimate are the LTQs and we decided to characterize the prior distribution of LTQs for each gene only by its mean and variance, leading to a Gaussian prior on LTQs. Note that this is equivalent to assuming a log-normal prior distribution over transcription quotients. In summary, the effective difference between negative binomial noise models and Sanity's noise model is that we convolve the Poisson distribution with a log-normal rather than with a gamma distribution. Apart from the theoretical motivation, we also prefer the log-normal model because it naturally arises under multiplicative noise (via the central limit theorem), and it naturally implements that the expected variance in expression scales as the square of the mean. However, it is possible that the choice of gamma versus log-normal may not make a large difference in practice.

4.7.4 The coefficient of variation and variance in log-expression are equal in the limit of small variations

For a random variable x with mean m and variance v, its coefficient of variation CV is defined through

$$CV^2 = \frac{v}{m^2}. \tag{4.101}$$

The variance of the logarithm of x is given by

$$\mathrm{var}\left(\log(x)\right) = \langle [\log(x) - \langle \log(x) \rangle]^2 \rangle. \tag{4.102}$$

If we write $x = m + \epsilon\sqrt{v}$, with ϵ a random variable with mean zero and variance 1, then we can write to first order

$$\log(x) = \log\left(m + \epsilon\sqrt{v}\right) \approx \log(m) + \epsilon\frac{\sqrt{v}}{m}, \tag{4.103}$$

which is accurate whenever $\sqrt{v}$ is significantly smaller than m, i.e. when the CV is small. We then find for the variance in $\log(x)$:

$$\mathrm{var}\left(\log(x)\right) \approx \langle \left[\epsilon\frac{\sqrt{v}}{m}\right]^2 \rangle = \frac{v}{m^2} = CV^2. \tag{4.104}$$

Thus, when the variations in gene expression are small relative to the mean, the CV-squared equals the variance in log-expression.

4.7.5 Using Euclidean distance to measure distances between cells

Once we have decided that we want to measure the gene expression state of a cell by a vector of log transcription quotients (LTQs), it immediately follows that the difference in the expression of gene g between cells c and $\tilde{c}$ is given by $d_{c\tilde{c}}(g) = \log(\alpha_{gc}) - \log(\alpha_{g\tilde{c}})$, which is the log fold-change of the expected expression levels between cells c and $\tilde{c}$. Note that, because $\log(\alpha_{gc}) = \mu_g + \delta_{gc}$ in terms of the mean LTQ μ_g and the log fold-changes δ_{gc} that Sanity reports, we can also write

$$d_{c\tilde{c}}(g) = \delta_{gc} - \delta_{g\tilde{c}}. \tag{4.105}$$

However, it is less obvious how to best combine the log fold-changes $d_{c\tilde{c}}(g)$ of each of the genes into a single distance $d_{c\tilde{c}}$ between the two cells. By far the most commonly used metric is to simply calculate the Euclidean distance between the vectors $\vec{\delta}_c$ and $\vec{\delta}_{\tilde{c}}$, i.e.

$$d^2_{c\tilde{c}} = \sum_g \left(\delta_{gc} - \delta_{g\tilde{c}}\right)^2, \tag{4.106}$$

and this is also how we decided to calculate distances between cells in this work.

Although we are not aware of any compelling arguments for favoring other distance metrics over this standard Euclidean metric, it is to some extent arbitrary to use the Euclidean distance metric. In particular, there is no reason to assume that the Euclidean distance $d_{c\tilde{c}}$ between two cells corresponds to some true biophysical difference in the expression states. However, there are several good mathematical arguments for using Euclidean distance. First, the Euclidean distance is a true distance metric in the sense that it satisfies the triangle inequality, i.e. for any triplet of cells a, b, c, the distance from a to c cannot be larger than the sum of the distance from a to b and the distance from b to c. Second, our usage of the Euclidean distance is consonant with characterizing the expression variability of each gene by its mean and variance. That is, the variance v_g of the expression of gene g across the cells is equal to the expectation of the squared-deviation $\langle \delta^2_{gc} \rangle$, which is directly related to the average squared distance between two randomly chosen cells, i.e.

$$\langle (\delta_{gc} - \delta_{g\tilde{c}})^2 \rangle = \frac{1}{C^2} \sum_{c,\tilde{c}=1}^{C} \left(\delta_{gc} - \delta_{g\tilde{c}}\right)^2 = 2v_g. \tag{4.107}$$

That is, the expected squared Euclidean distance between two randomly chosen cells is twice the sum of the variances v_g, summed over all genes. Thus, when using Euclidean distances the distances between cells are directly related to the variances in expression across cells.

In addition, in contrast to most alternative metrics, the Euclidean distance also has the advantage that it is invariant not only under translation but under arbitrary continuous rotations of the vectors. Therefore, the distances between cells are

invariant to the set of orthonormal base vectors used to represent expression states. In principal component analysis one searches for lower dimensional subspaces that capture most of the variance in the data and, when using the Euclidean distance as a distance metric, this is equivalent to finding the lower dimensional subspace for which the averaged squared distance between the cells is largest. Finally, note also that, if all cells have approximately the same total squared distance from the global average expression profile, i.e. when

$$\sum_g \delta_{gc}^2 \approx V = \sum_g v_g \; \forall c, \tag{4.108}$$

then there is also a monotonic relationship between the squared distance between two cells, and the Pearson correlation of their expression profiles, i.e.

$$d_{c,\tilde{c}} = \sum_g \left(\delta_{gc} - \delta_{g\tilde{c}}\right)^2 \approx 2V \left(1 - r(\vec{\delta}_c, \vec{\delta}_{\tilde{c}})\right), \tag{4.109}$$

with $r(\vec{\delta}_c, \vec{\delta}_{\tilde{c}})$ the Pearson correlation of the two vectors (i.e. their inproduct normalized by the product of their lengths). Although none of this shows that usage of Euclidean distances is required, it does argue that it is a natural choice.

4.7.6 Correcting for batch effects with Sanity

The term batch effect is used to refer to a wide array of uncontrolled variations in experiments of both technical and biological nature, and some normalization methods also specifically aim to correct for batch effects. In our opinion, normalization methods should not conflate true biological differences in gene expression (for example due to different genotypes or experimental conditions) from differences that result from technical variation in the experimental protocols. Normalizing away differences that correspond to true biological differences may well remove evidence of important biological effects. We thus feel that normalization methods should only aim to correct for technical variation and we here briefly discuss what types of technical variations Sanity can and cannot correct for.

First, we note that there are types of experimental variability that no normalization method (that we are aware of) corrects for. For example, the efficiency in extraction of cells of different types from the tissue or system under investigation may vary across protocols. Such biases will crucially affect the way cells from a given sample will appear to be distributed in gene expression space. However, since normalization methods, including ours, only quantify the expression states of the cells that *were* captured, they cannot correct for such cell capture biases.

There are also technical variations that Sanity automatically corrects for. First, since Sanity defines the gene expression state of a cell in terms of the vector $\vec{\alpha}_c$ of transcription quotients, the method naturally corrects for variations in the total

UMI count per cell. This includes variations in total UMI count both due to variations in cell size as well as due to technical variation in the efficiency of mRNA capture and sequencing. In particular, Sanity automatically corrects for arbitrary variations in the total transcription activities A_c and the capture and sequence probabilities p_c across cells.

Sanity assumes that, in a given cell c, each mRNA has the *same* probability p_c to be sequenced. However, it is plausible that in reality mRNAs for different genes have different probabilities to be captured and sequenced. Note that, if the capture and sequence probability p_{gc} of gene g in cell c were to vary in an arbitrary way across both genes and cells, it would be impossible to disentangle fluctuations in gene expression from fluctuations in p_{gc}. However, at least for the cells within one experiment, it seems reasonable to assume that the relative capture and sequencing efficiencies of different genes are the same. Under that assumption, we would have that the capture and sequence probability for gene g in cell c is a product of a cell-specific factor and a gene specific factor, i.e

$$p_{gc} = p_c q_g. \tag{4.110}$$

Note that in this model the relative probability of capturing an mRNA of gene g in two cells c and c' is $p_c/p_{c'}$ independent of the gene, and the relative probability of capturing two different genes g and g' is $q_g/q_{g'}$ independent of the cell. That is, in this model different genes may have different propensities q_g to be captured, and capture efficiency p_c may vary across cells, but the relative propensities across genes stays the same across cells within one experiment.

If we define the rescaled average LTQs across genes $\beta_g = q_g \alpha_g$, then all the calculations that we performed above go through in the exact same way by simply replacing α_g with β_g everywhere in the equations. In addition, note that equation (4.110) above is invariant under rescaling all the q_g by some constant X if we at the same time rescale all p_c by $1/X$. Thus, without loss of generality we can freely pick the normalization of the q_g and a natural choice is to demand that

$$\sum_g \beta_g = \sum_g q_g \alpha_g = 1. \tag{4.111}$$

This way, the β_g are still normalized so that they sum to 1.

We can then simply reinterpret the estimated average LTQs μ_g as the sum of the true biological average LTQ $\log(\alpha_g)$ and shift $\log(q_g)$ due to technical capture-and-sequence bias, i.e.

$$\mu_g = \log(\beta_g) = \log(\alpha_g) + \log(q_g). \tag{4.112}$$

In other words, the mean LTQs that we estimate for each gene are the sum of the true mean LTQ plus an unknown correction term due to (unknown) technical bias in the efficiency and capture of different genes.

Importantly, while the biases q_g are likely roughly constant across cells within one experiment, and maybe even across multiple experiments that use the same protocols, we can easily imagine that the q_g will change when the scRNA-seq protocol is changed. However, it should now be clear how one can correct for such batch effects. When combining data from different 'batches' of scRNA-seq experiments, we can simply run Sanity separately on the UMI counts of each of the batches b. The estimated log fold-changes δ_{gc} and their error bars ϵ_{gc} do not depend on variations in the q_g and can be combined directly from the different batches. However, for each batch b and each gene g we will get a separate estimated mean LTQ μ_{gb}. We can then define a final mean LTQ μ_g for each gene by simply averaging the μ_{gb} across the batches, i.e.

$$\mu_g = \frac{1}{B} \sum_b \mu_{gb} = \log(\alpha_g) + \langle q_g \rangle, \tag{4.113}$$

where B is the number of batches and $\langle q_g \rangle$ is the bias of gene g averaged over all batches.

4.7.7 Sanity's use of a Gaussian prior on LTQs does not preclude it from correctly identifying non-Gaussian expression distributions

Since Sanity uses a Gaussian distribution for the prior probabilities $P(\delta_{gc}|v_g)$ of the log fold-changes across cells, one could suspect that this makes it impossible for Sanity to correctly identify the expression distributions for genes whose expression is bimodally distributed, or have some other non-Gaussian distribution. However, this is not the case. The crucial point to note is that the Gaussian distribution is only a *prior* distribution used to estimate the total variance in LTQ of each gene. The estimated log fold-changes δ_{gc} that Sanity infers do not only depend on this prior, they are in fact mostly driven by the observed UMI counts n_{gc}. In particular, if these UMI counts show clear evidence of bimodal expression, Sanity will infer bimodally distributed log fold-changes δ_{gc}.

To explicitly demonstrate this we created a simulated dataset with 2000 cells whose total UMI counts N_c were drawn from a lognormal distribution with mean 10^4 and standard deviation 925 (which roughly matches the distribution of total counts in real data) and containing 4 genes. The log transcription activities of these genes were chosen as follows

1. Gene 1 had Gaussian distributed log transcription activities with mean 1 and variance 1.

2. Gene 2 had bimodally distributed log transcription activities consisting of an equal mixture of two Gaussians, one with mean 1 and variance 0.1 and one with mean log(20) and variance 0.1.

3. Gene 3 had uniformly distributed log transcription activities in the range $[0, 4]$.

4. Gene 4 was a dummy gene that was used to set the total transcription activity in each cell c to the desired value N_c.

We first sample the log transcription activities $\log(a_{gc})$ for the first 3 genes across all cells. Then, for each cell c the UMI counts of the first 3 genes are sampled from a Poisson distribution with mean equal to the value of a_{gc} that was sampled. Finally, the UMI count for the 4th gene is sampled from a Poisson distribution with mean $N_c - \sum_{g=1}^{3} n_{gc}$, i.e. so as to ensure the expected total count is N_c. Note that the true LTQ of genes one through three are given by $\alpha_{gc} = \log(a_{gc}/N_c)$. We then ran Sanity on this simulated dataset and subsequently ignored the results for the dummy gene 4.

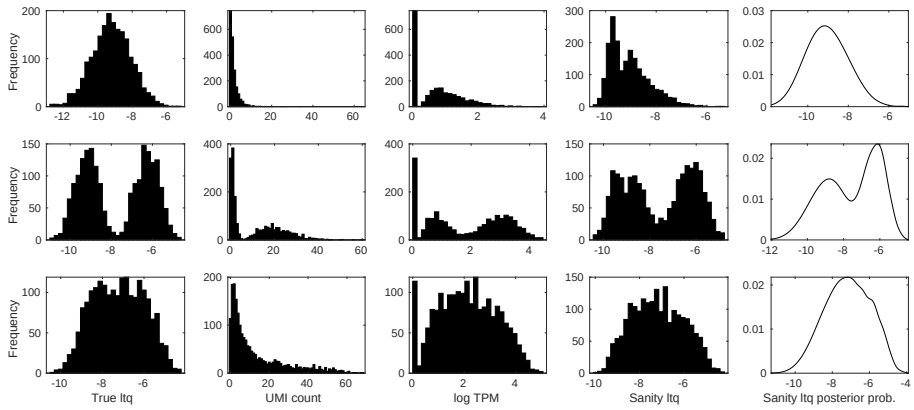


Figure 4.28: Sanity's use of a Gaussian prior on LTQs does not preclude it from correctly identifying bimodal and other non-Gaussian expression distributions. We simulated data from three genes whose true LTQs are either Gaussian distributed (top row), bimodally distributed (middle row), or uniformly distributed (bottom row). Each column of panels shows, from left to right: the histograms of true LTQs across cells, the histogram of raw UMI counts across cells, the histogram of normalized log-expression levels using the simple TPM normalization, the histogram of LTQs values that Sanity inferred, and the inferred overall distribution of LTQs obtained by taking the mixture of posterior distributions for the LTQ in each cell.

Suppl. Fig. 4.28 shows, for each of the three genes, the distribution of true LTQs, the distribution of raw UMI counts, the histogram of log-expression values when using the simple TPM normalization method, the histogram of the LTQ values $\mu_g + \delta^*_{gc}$ that Sanity estimates, and finally, the overall distribution of LTQ values that Sanity infers. The distribution in this rightmost column of panels is obtained by taking the estimated LTQs $\mu_g + \delta^*_{gc}$ and their error bars ϵ_{gc} and

calculating the mixture of Gaussians with means $\mu_g + \delta^*_{gc}$ and variances ϵ^2_{gc} across all cells.

For gene 1, whose LTQs are Gaussian distributed and whose average expression is fairly low, the raw UMI counts are peaked at zero and when using the simple TPM method, this leads to a clearly bimodal distribution of log-expression, with cells with zero UMIs clearly separating from cells with one or more UMIs (Fig. 4.28, top row, second and third columns from left). The histogram of LTQ values reported by Sanity is already much closer to the true distribution of LTQs (Fig. 4.28, top row, second column from the right). However, the inferred LTQ values for cells in which there are zero UMI counts still show up as a second mode to the left of the true mode of the distribution. Note that, at the same time, the tail of LTQ values is truncated relative to the true distribution (Fig. 4.28, top row, leftmost column). What is happening is that for most cells for which the true LTQ is below -9.5, the observed UMI count is zero. Since the only information we have about the true LTQ $\log(\alpha_{gc})$ in a cell is the observed UMI count n_{gc}, the total count N_c, and an estimate of the total mean μ_g and variance v_g in LTQ, the estimated δ_{gc} for cells with zero UMI are all very similar, leading to a peak in LTQ around -9.5 in the fourth panel from the left. However, for each of these cells Sanity also reports a substantial error bar ϵ_{gc} for the estimated δ_{gc}, which indicates that the true LTQ values in those cells could be anywhere in the range $[-12, -9]$ or so. If we take these error bars into account and reconstruct an overall distribution of LTQs by averaging over all cells, i.e. we define

$$P(\delta_g) = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{\sqrt{2\pi}\epsilon_{gc}} \exp\left[-\frac{(\delta_{gc} - \delta^*_{gc})^2}{2\epsilon^2_{gc}}\right], \tag{4.114}$$

with C the total number of cells, then we see that this distribution $P(\delta_g)$ is in fact very close to the true distribution of LTQs (Fig. 4.28, top row, rightmost panel).

For gene 2 the true distribution of LTQs is bimodal and this bimodality is even seen in the distribution of raw UMI counts (Fig. 4.28, middle row, first two columns from the left). The frequent occurrence of cells with zero UMI count still leads to an extra mode for the TPM normalized log-expression values. Most importantly, the LTQ values that Sanity estimates have a bimodal distribution that quite closely approximates the true bimodal distribution of LTQ values, showing that Sanity's Gaussian prior does not preclude the algorithm from correctly inferring bimodal distributions of LTQ values. Note, however, that because the Gaussian prior assigns highest probability to values in between the two modes, the modes of the inferred LTQ values are slightly less well separated than the true LTQs, as is most clearly visible when plotting the distribution of LTQ values, as calculated using equation (4.114) (Fig. 4.28, middle row, rightmost column).

Finally, for gene 3 with a wider and more uniform distribution of LTQs, Sanity also correctly infers a broad distribution of LTQ values that is approximately flat over the same range as the true LTQ values (Fig. 4.28, bottom row).

4.7.8 Estimated variances v_g for very lowly expressed genes

As shown in Suppl. Figs. 4.10 and 4.11, when the absolute expression levels of a gene are very low, Sanity can often estimate a variance of true LTQs v_g that is much lower than the true variance, and one might wonder if this is not a systematic error. However, as we show here, at such low expression levels the variance in LTQs is simple not detectable from the data. Conceptually, one can easily have that a gene's LTQs vary over a wide range but with such low absolute values that the expected number of UMIs in each cell is still below one. In that case, all that would be observed in the data is a single UMI in a few cells, and zero UMI in all other cells. From such data it is impossible to infer the true amount of variability in LTQs because all this variation occurs below the detection limit. We illustrate this with simulated data.

Using the same simulation scheme as described in the previous section, we simulated the UMI counts in $C = 1000$ cells, with total UMI counts N_c distributed around 10^4, for 3 genes with the following true distributions of log transcription activities

1. Gene 1 has a mean log transcription activity of $\mu_g = -1.9$ with no variation at all, i.e. the log transcription activity is the same in every cell.

2. Gene 2 has Gaussian distributed log transcription activities with mean $\mu_g = -2$ and variance $v_g = 0.1$

3. Gene 3 has Gaussian distributed log transcription activities with mean $\mu_g = -2.8$ and variance $v_g = 1$.

Thus, the 3 genes have very different true variation in LTQ, with gene 1 not varying at all, gene 2 showing moderate variation, and gene 3 showing fairly large variations in LTQ. The true histograms of log transcription activity for these 3 genes are shown in Fig. 4.29, top row.

Note that the mean expression levels of these genes were chosen on purpose to be so low that, for about 90% of cells no UMIs are observed, and the distributions of UMI counts are very similar across the 3 genes (Fig. 4.29, bottom left). That is, all three genes have zero UMI in about 90% of the cells, one UMI in about 100 cells, and two UMI in around 10 cells. In fact, the distribution of UMI counts for genes 1 and 2 is almost exactly the same, i.e. with the same number of cells for which 1 or 2 UMIs are observed. The only difference is that there is *one* cell in which gene 2 has 3 UMIs, whereas there is no cell with 3 UMIs for gene 1. The highly variable gene 3 also has an almost exactly equal distribution of UMI counts, differing only by having slightly more cells with 2 UMIs and a single cell with 4 UMIs. Thus, the raw data for these three genes are almost identical.

When Sanity is run on this data, it returns the posterior distributions over the true variances v_g shown in the middle row of Fig. 4.29. For gene 1, Sanity

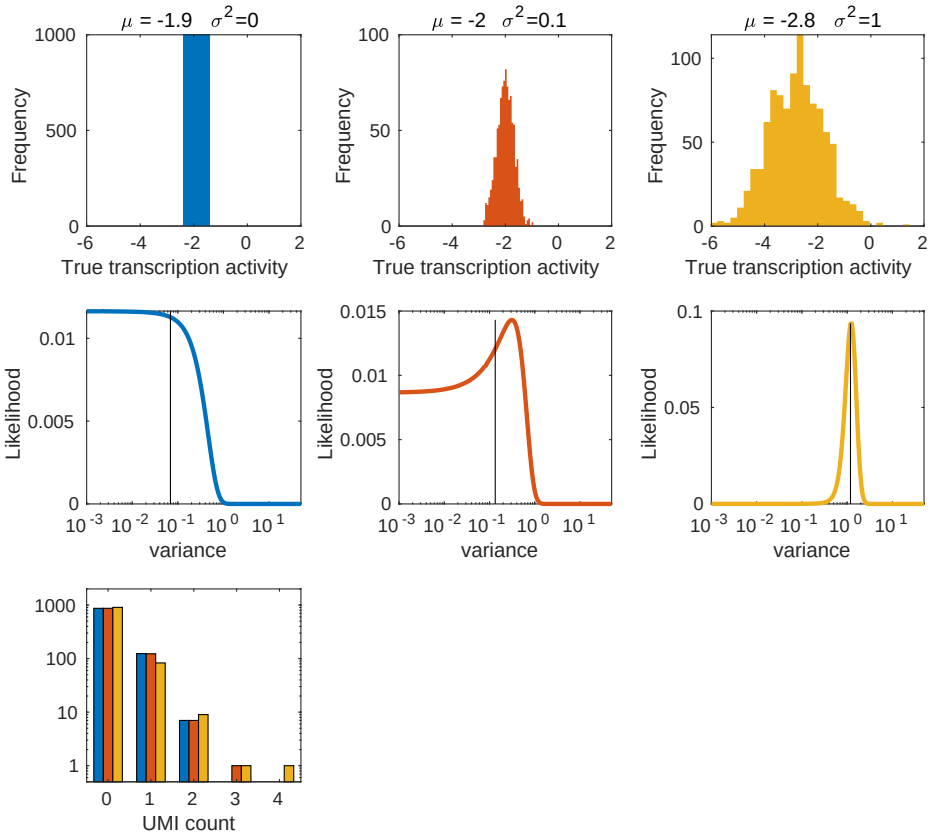


Figure 4.29: Estimates of true variance in LTQ for low expressed genes. We simulated data for three genes whose LTQs are Gaussian distributed with variance $v_g = 0$ (left), $v_g = 0.1$ (middle), and $v_g = 1$ (right), and with means μ_g chosen such that each gene is expected to have zero UMIs in approximately 90% of the cells. The top row shows the histogram of true log transcription activities for each of the genes across the $C = 1000$ cells. The bottom left panel shows the distribution of raw UMI counts for the three genes, which are almost identical. The panels in the middle row show the posterior distributions $P(v_g|\vec{n}_g)$ for each of the genes, with the vertical line showing the expected value $\langle v_g \rangle$.

correctly infers that the most likely value of the variance v_g is zero, although values of the variance v_g as high as 0.5 cannot be excluded. The estimate that Sanity returns is given by an average over this posterior distribution, which is about 0.07 for gene 1. For gene 2, we see that the single cell with UMI count of 3 already noticeably alters the posterior distribution over v_g. While the most likely value occurs around $v_g = 0.3$, there is a constant tail to the left showing that the data is

almost equally consistent with arbitrarily low values of v_g. Thus, given the data it cannot be excluded that this gene's LTQ does not vary at all. Taking the average over the posterior distribution of gene 2 leads to an estimate v_g slightly above the true value of 0.1. Finally, for gene 3, in spite of the almost identical distribution of UMI counts, Sanity's posterior distribution is highly peaked around the true value of $v_g = 1$.

In summary, when genes are very low expressed, distributions of LTQs with very different variances can lead to distributions of UMI counts that are almost identical. However, even when these distributions of UMI counts are almost identical, Sanity is exquisitely sensitive to the precise distribution of UMI counts, and can detect true variations in LTQ even from a few extra cells with one or two extra UMIs. Consequently, when Sanity infers that there is no evidence in the data for true variation in LTQs, it really means there is no information whatsoever to support that the gene's LTQ varies across cells. Although this does not guarantee that the gene does not vary in LTQ across cells, it means that this variation is entirely below the limit of detection, so that it is impossible to tell in which cells the gene is higher or lower expressed. Given that, it is appropriate for Sanity to default to assigning no variability in the LTQ at all, i.e. to predict a low value for v_g.

4.7.9 The fraction of genes for which expression levels can be accurately estimated depends strongly on coverage

Fig. 4.3 in the main text showed that, because of the inherent Poisson noise, accurate estimates of gene expression can only be guaranteed for genes with at least 1 UMI per cell on average. Consequently, the number of genes for which expression levels can be accurately estimated depends on the capture efficiency and depth of sequencing, as well as on the distribution of absolute expression levels across genes.

As shown in Fig. 4.30, the distributions of mean expression levels μ_g across genes are very similar for the datasets that we analyzed here. The most common value of the mean LTQ μ_g lies between 10^{-5} and 10^{-4}, and the frequency drops must faster toward higher μ_g than toward lower μ_g. Consequently, genes with $\mu_g > 0.001$ are extremely rare.

Given that the median number of UMI per cell N_c ranges from about 2000 to $10'000$ for the datasets we analyzed here, this means that for the large majority of genes the expected number of UMI per cell is less than 1. We pooled data from all the distributions shown in Fig. 4.30 to obtain a single distribution of mean expression μ_g across genes and then determined the reverse cumulative distribution of the average number of UMI per cell, assuming different total UMI counts N_c ranging from $N_c = 1000$ to $N_c = 100'000$ (i.e. tenfold larger than achieved with current scRNA-seq protocols). These reverse cumulative distributions are shown

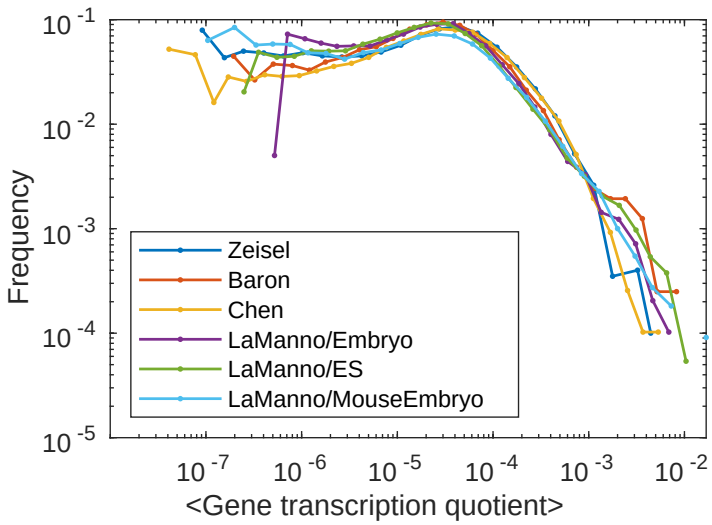


Figure 4.30: Distributions of the mean LTQs μ_g across all genes, for the datasets analyzed in this work (colors, see legend). Each curve corresponds to the distribution of mean LTQs μ_g as inferred by Sanity with μ_g shown on the horizontal axis and the probability density shown on the vertical axis. Note that the vertical axis is shown on a logarithmic scale.

in the left panel of Fig. 4.31.

From these reverse cumulative distributions we can calculate the fraction of genes that have at least 1 UMI per cell on average, as a function of the total UMI count N_c (Fig. 4.31, right panel). Given that the median UMI counts per cell ranged from 2000 to 10′000 for the datasets analyzed here, we see that even for the datasets with the deepest coverage of $N_c \approx 10^4$, only about 10% of genes have at least 1 UMI per cell on average. Thus, with current depth of coverage we can only get accurate estimates of gene expression patterns for the 10% highest expressed genes. This means that, if certain cell states only differ in the expression states of more lowly expressed genes, it would be very difficult if not impossible to distinguish between these with current depth of coverage. Most importantly, however, this also means that if we could increase the capture and sequencing probability, we could substantially increase the fraction of genes for which gene expression can be estimated accurately. In particular, if we increase the capture and sequencing probability by a factor 3, we would be able to get accurate estimates for approximately 25% of the genes, and if we could capture 10^5 UMI per cell, we could get accurate estimates for over 40% of the genes. Finally, for comparison, the total number of mRNAs in mammalian cells is likely in the range $10^5 - 10^6$ [Islam et al., 2014].

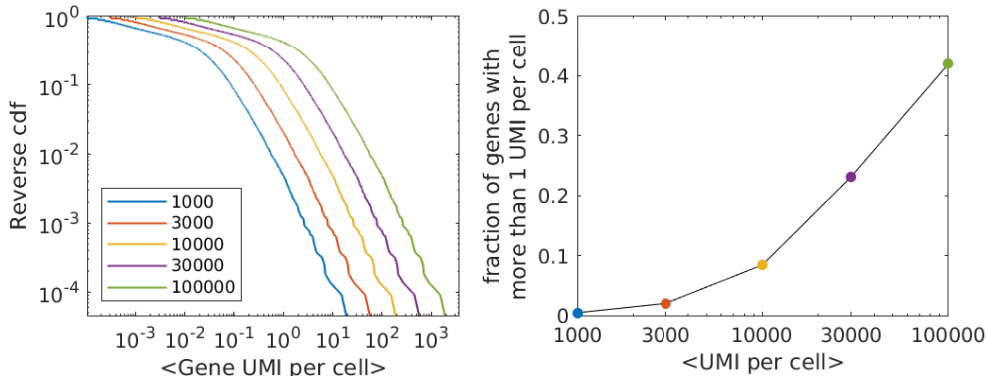


Figure 4.31: **Left panel**: Reverse cumulative distributions of the average number $\langle n \rangle$ of UMI per cell across genes (horizontal axis), i.e. the fraction of genes with at least $\langle n \rangle$ UMI per cell on average, for 5 different values of the total number of UMI per cell N_c (colors, see legend). Both axes are shown on logarithmic scales. **Right panel**: Fraction of the genes that have at least 1 UMI per cell on average (vertical axis), as a function of the total number of UMI per cell N_c (horizontal axis). The colored dots correspond to the N_c values used for the corresponding curves in the left panel. The horizontal axis is shown on a logarithmic scale.

Chapter 5

Realizing Waddington's metaphor: Inferring regulatory landscapes from single-cell gene expression data

Jérémie Breda[1,2], Tanzila Mukhtar[3], Mikhail Pachkov[1,2], Verdon Taylor[3], Mihaela Zavolan[1,2], Erik van Nimwegen[1,2]

1. Biozentrum, University of Basel, Basel, Switzerland
2. Swiss Institute of Bioinformatics, Basel, Switzerland
3. Department of Biomedicine, University of Basel, Basel, Switzerland

Unpublished work

5.1 Introduction

The cellular complexity of higher eukaryotes originates from a single egg cell that divides and differentiate into a myriad of cells having identical genomic information but differing by the expressed segments of the genome. Different regulatory mechanisms control DNA usage at the level of chromatin remodelling, transcription, RNA splicing and processing, export to cytoplasm, translation into proteins and ultimately degradation [Orphanides and Reinberg, 2002]. Every regulator of gene expression is itself a product of gene expression, creating the gene regulatory networks (GRNs) [Davidson and Erwin, 2006, Davidson, 2010]. Whereas noise in gene expression could naively be viewed as a nuisance, more recent work started to reveal other interesting aspects. For example, the coupling between a target gene and a noisy regulator could be viewed as a rudimentary form of gene regulation [Wolf et al., 2015]. As gene expression is known for being inherently stochastic [Mcadams and Arkin, 1997, Elowitz et al., 2002, Paulsson, 2005, Raj and van Oudenaarden, 2008], we expect regulatory interactions to define a probability density of cell states in the high dimensional space of gene expression. That is, GRNs engender a function defined in gene regulatory space analogous to a potential energy which constrains the stochastic fluctuations as well as the continuous changes in the gene expression state of cells. Such a system, involving a large number of particles with high degree of freedom, with an inherently stochastic nature and under the influence of an energy potential, have been deeply described by statistical physics, and from this premise we conclude that the distribution of cells in gene expression space subject to regulatory interactions, will follow the maximum entropy distribution given only the energy function generated by GRNs, known as the Boltzmann distribution. Under this description, gene regulatory interactions can be observed from noisy fluctuations across similar cells, and cellular process in which the cell state varies (*e.g.* differentiation, development, responses to stimuli or malignant transformation) are defined as continuous regions in gene expression of high probability density.

On the task of inferring gene regulatory interaction from gene expression data, the MARA model has brought valuable and validated insights about gene regulation in various systems since its initial release [Gruber et al., 2014, Grunin et al., 2016, Yan et al., 2016, Dimitrova et al., 2017, Yeung et al., 2018, Tauran et al., 2019, Danoy et al., 2019], but was primarily built for bulk RNA sequencing. To account for the noise specific to single-cell RNA sequencing (scRNA-seq) [Grün et al., 2014], we developed a Bayesian model [Breda et al., 2021], such that gene regulatory interaction can be inferred from biological signal. The idea of a GRN creating a potential surface constraining the regions of gene expression that cells can reach, is closely related to the idea of an epigenetic landscape as imagined by Conrad Hal Waddington in 1957 [Waddington, 1957].

Cell differentiation is a central question of biology and is still partially under-

stood today. In 1957, Waddington introduced the idea of an epigenetic landscape driving cell differentiation [Waddington, 1957]. In this analogy, a stem cell evolves similarly to marbles rolling down a surface. The shape of the surface creates multiple path and branching, that the marbles can follow, and end up in one of the stable minima of the surface. Similarly, a stem cell would evolve through gene expression space to end up in a stable state called mature cell type (See Figure 5.1). As a consequence of discoveries in cell reprogramming, there has been a surge in activity in this area, and Waddington's representation remains one of the most popular metaphors employed in thinking about how regulatory networks control cell state and identity. However, it remains unclear how Waddington's landscape picture can be taken beyond an attractive metaphor.

Here we propose a computational framework, that, by combining a new noise model for scRNAseq [Breda et al., 2021] with modelling of transcriptional states in terms of the activities of transcriptional regulators [Balwierz et al., 2014], reconstructs an explicit epigenetic landscape that can be used to identify stable cellular states and the regulators that control their stability, and identify developmental path.

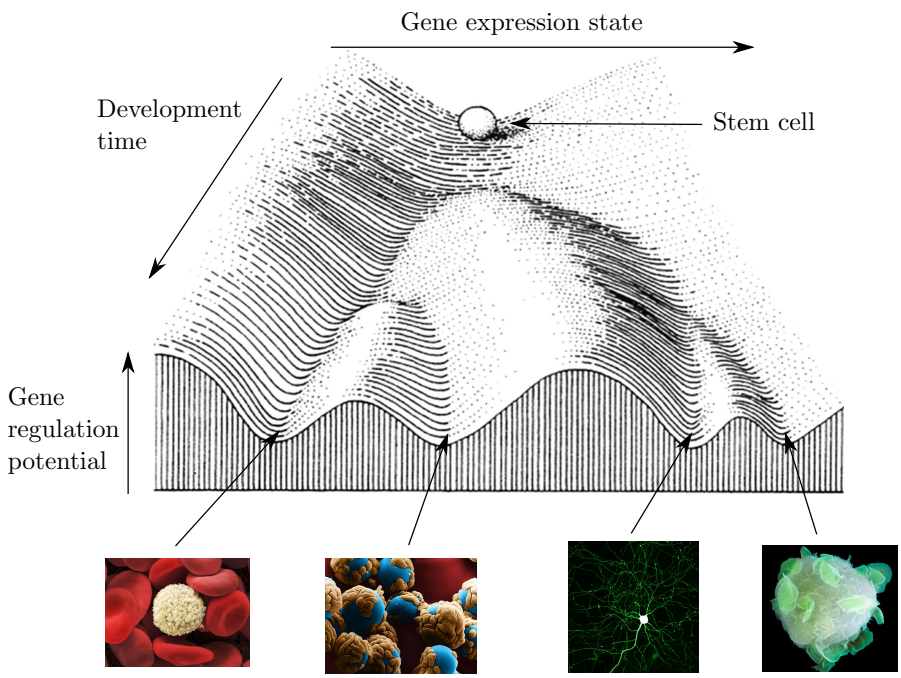


Figure 5.1: Waddington's epigenetic landscape. A stem cell moves through gene expression space along developmental time to end up in a stable cell type state. (From [Waddington, 1957])

5.2 Model and methods

To infer a regulatory potential energy from gene expression, we use the observed fluctuations in gene expression across a population of cells from the same type. From the point of view of physics every system can be characterized by an energy function. And since the studied system shows inherent stochasticity together with large degree of freedom, we can use the mathematical framework of statistical physics. That is, the probability of finding a system in a state s, knowing only the energy of states $E(s)$, follows the maximal entropy distribution, known as Boltzmann distribution:

$$P(s) = e^{-\beta E(s)} \tag{5.1}$$

As scRNA-seq technology measures of distribution of cell in gene expression space, we can use this framework to infer an energy function $E(s)$. However, a scRNA-seq experiment really measures a number of transcript for every gene. To obtain a density in gene expression $P(s)$ we developed the method Sanity presented previously in this work (chapter 4, [Breda et al., 2021])

In his 1957 book, The Strategy of the Genes [Waddington, 1957], Waddington imagined a "Complexe system of interactions underlying the epigenetic landscape" controlled by the "chemical tendencies which genes produce" as he illustrated on Figure 5.2. Indeed, regulators are driving gene expression. The major regulators of gene expression being transcription factor we used a model previously developed in the group of Erik van Nimwegen called MARA (for Motif Activity Response Analysis) [Balwierz et al., 2014]. MARA allow us to infer the landscape in the space of regulators rather than genes. It has a much lower dimension and the landscape we infer directly reveals gene regulators interactions. Furthermore, The fact that each regulator activity can be inferred from the expression of many genes considerably reduces the noise.

Briefly, MARA for each gene identifies binding motifs of transcription factors on the promoter region and miRNA on the 3'-UTR of the transcripts. The model then interprets the measured gene expression of promoter p in cell c E_{pc} as a linear combination of the number of binding sites on promoter p for each motif m N_{pm} multiplied by the activity of the corresponding motifs m in cell c A_{mc}:

$$E_{pc} = \sum_m N_{pm} A_{mc} \tag{5.2}$$

Let $\mathbf{A}^c \in \mathbb{R}^M$ the maximum a posteriori esitmates of the activity vector of cell c in motif activity space, with M the number of motifs. Given this cell only, the probability of $\mathbf{A}$ is[1] :

[1] from equation (8) of *ISMARA: Automated modeling of genomic signals as a democracy of regulatory motifs* supplementary material

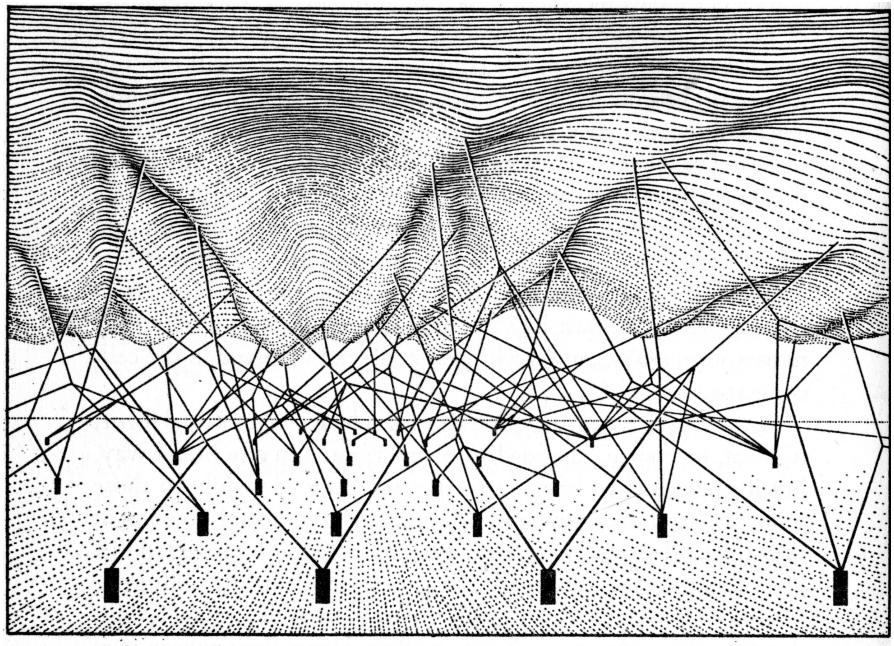


FIGURE 5
The complex system of interactions underlying the epigenetic landscape.

Figure 5.2: Waddington's schematic view of regulatory interactions underlying his epigenetic landscape and responsible for its shape. (From [Waddington, 1957])

$$P(\mathbf{A}|c, N, \sigma) \propto \sigma^{-P} \exp\left(\frac{\sum_{mn}(A_m - A_m^c)W_{mn}(A_n - A_n^c) + \chi_c^2}{2\sigma}\right) \tag{5.3}$$

with $\chi_c^2 = \sum_p (E_{pc} - \sum_m N_{pm} A_{mc})^2$.

Using the scale-invariant prior $P(\sigma) \propto \frac{1}{\sigma}$ and marginalizing over the unknown σ, we obtain the probability of activity $\mathbf{A}$.

$$\begin{aligned} P(\mathbf{A}|c, N) &= \int_0^\infty d\sigma P(\sigma) P(\mathbf{A}|c, N, \sigma) && (5.4) \\ &\propto \frac{\Gamma\left(\frac{P}{2}\right)}{[\sum_{mn}(A_m = A_m^c)W_{mn}(A_n - A_n^c) + \chi_c^2]^{\frac{P}{2}}} && (5.5) \\ &\propto \exp\left[-\frac{P\sum_{mn}(A_m = A_m^c)W_{mn}(A_n - A_n^c)}{2\chi_c^2}\right] && (5.6) \end{aligned}$$

Introducing a global temperature parameter β, we have

$$P(\mathbf{A}|c) = \exp\left(-\frac{\beta}{2}(\mathbf{A} - \mathbf{A}^c)^\top \frac{P\mathbf{W}}{\chi_c^2}(\mathbf{A} - \mathbf{A}^c)\right) \tag{5.7}$$

As $\mathbf{W}$ is real and symmetric, it has an orthonormal basis S so that $W = S^\top \Lambda S$, where Λ is a diagonal matrix of eigenvalues λ_i.

So we can generally write :

$$\sum_{m,n}(A_m - A_m^c)W_{mn}(A_n - A_n^c) = \sum_i \lambda_i (X_i - X_i^c)^2 = \sum_i (Y_i - Y_i^c)^2 \tag{5.8}$$

where $X_i = S_{in}A_n$ and $Y_i = \sqrt{\lambda_i}X_i$.

In this basis the probability of state $\mathbf{Y}$ is given by

$$P(\mathbf{Y}|c) = \exp\left(-\frac{\beta P}{2\chi_c^2}\sum_i (Y_i - Y_i^c)\right). \tag{5.9}$$

Let's define $w_c = \frac{P}{\chi_c^2}$.

Taking the maximal entropy distribution of cell motif activity state assuming only its mean activity, we use the Boltzann distribution to define the energy function

$$E(\mathbf{Y}) = -\frac{1}{\beta}\log\left[\sum_c P(\mathbf{Y}|c)\right] = -\frac{1}{\beta}\log\left[\sum_c \exp\left(-\frac{\beta}{2}w_c\sum_i (Y_i - Y_i^c)^2\right)\right]. \tag{5.10}$$

A minima of $E(\mathbf{Y})$ must obey the equation

$$\frac{\partial E(\mathbf{Y})}{\partial Y_i} = \frac{\sum_c P(\mathbf{Y}|c)w_s(Y_i - Y_i^c)}{\sum_c P(\mathbf{Y}|c)} = 0 \ \forall i \tag{5.11}$$

implying

$$\sum_c P(c|\mathbf{Y})w_c Y_i = \sum_c P(c|\mathbf{Y})w_c Y_i^c \tag{5.12}$$

where we define

$$P(c|\mathbf{Y}) = \frac{P(\mathbf{Y}|c)}{\sum_c P(\mathbf{Y}|c)}. \tag{5.13}$$

A minma $\mathbf{Y}^*$ must then satisfy

$$\mathbf{Y}^* = \frac{\sum_c w_c \mathbf{Y}^c P(c|\mathbf{Y})}{\sum_c w_c P(c|\mathbf{Y})}. \tag{5.14}$$

The second derivative at $\mathbf{Y}^*$ is given by

$$\frac{\partial^2 E(\mathbf{Y})}{\partial Y_i \partial Y_j} = \sum_c w_c P(c|\mathbf{Y}^*)\left[\delta_{ij} - w_c(Y_i^c - Y_i^*)(Y_j^c - Y_j^*)\right] \tag{5.15}$$

Because the first term is proportional to the identity matrix, the eigenvectors of the full Hessian are also eigenvectors of this sort of covariance-matrix

$$C_{ij}^* = -\sum_c w_c P(c|\mathbf{Y}^*)w_c^2(Y_i^c - Y_i^*)(Y_j^c - Y_j^*) \tag{5.16}$$

5.3 Results

The derivations in section 5.2 show how to derive a potential function defined in regulatory motif activity space from scRNA-seq data. We use two data sets to apply this framework on two questions; inferring cell sub-types and developmental trajectories.

5.3.1 Mature cells of human pancreas

Equation (5.10) defines an energy function in motif activity space and its minima, characterized by equation (5.14), represent an intuitive way of defining cell types as local regions of high cell density. To test this hypothesis we used a published set of human pancreas cells from 4 human donor [Muraro et al., 2016]. Figure 5.3.a shows all 2'298 cells projected on the first 3 principal components, capturing 46% of the total variance.

Equation (5.14) is a condition each local minimum must satisfy. We can use it as an expectation maximization (EM) equation: starting from any vector $\mathbf{Y}$, we iterate (5.14) until reaching a stable point $\mathbf{Y}^*$. As each local minimum is located in a region of relatively high density of cells, we argue that starting the EM recursion from each cell guarantees to find every existing minimum. Each found minimum is defined as a cell type and each cell is assigned to the cell type of the local minimum reached by the EM recursion that started from that cell. This procedure is illustrated on Figure 5.3.b with the inferred landscape projected on the first 2 principal components defined by the cells. Equation (5.9) shows that the landscape depends on the temperature parameter β. This unique parameters scales the contribution of each cell to the overall landscape and determines its number of minima. Figure 5.3.c, shows that for small β, there is only one minumum containing all cells whereas high β implies that no minima with at least 2 cells is found, meaning that each cell is in its own minimum. Hence, only a restricted range of values of β between 0.02 and 0.12 produces some cell types. We chose the value of β maximizing the number of minima with at least 10 cells. Applying this method we find 7 minima. To assign the minima to different cell types of pancreatic cells, we used the set of marker genes associated with each cell type that have been found in the publication [Muraro et al., 2016]. As the publication does not provide cell types annotation but a list of marker genes per identified cell types, we used this list of marker genes sorted by significance to compared our cell types using genes differentially expressed in each type. The identified cell types are annotated on figure 5.3.e on all cells colored by type and projected on the first 3 principal components.

As those cell types define regions in the space of regulatory motif activity, we asked which regulator best distinguish the cell types. The 15 regulators which best distinguish the cell types are shows as projected axis on Figure 5.3.d, and some

examples of regulatory motif activity distributions are showed on figure 5.3.e. From the literature, we could corroborate those factors as important regulators of pancreatic cells [Xuan and Sussel, 2016, Nishimura et al., 2006, Maity et al., 2018, Dooley et al., 2016, Algül et al., 2007, Martin et al., 2015, Ait-Lounis et al., 2007, Cebola et al., 2015, Aghdassi et al., 2012, Qian et al., 2017].

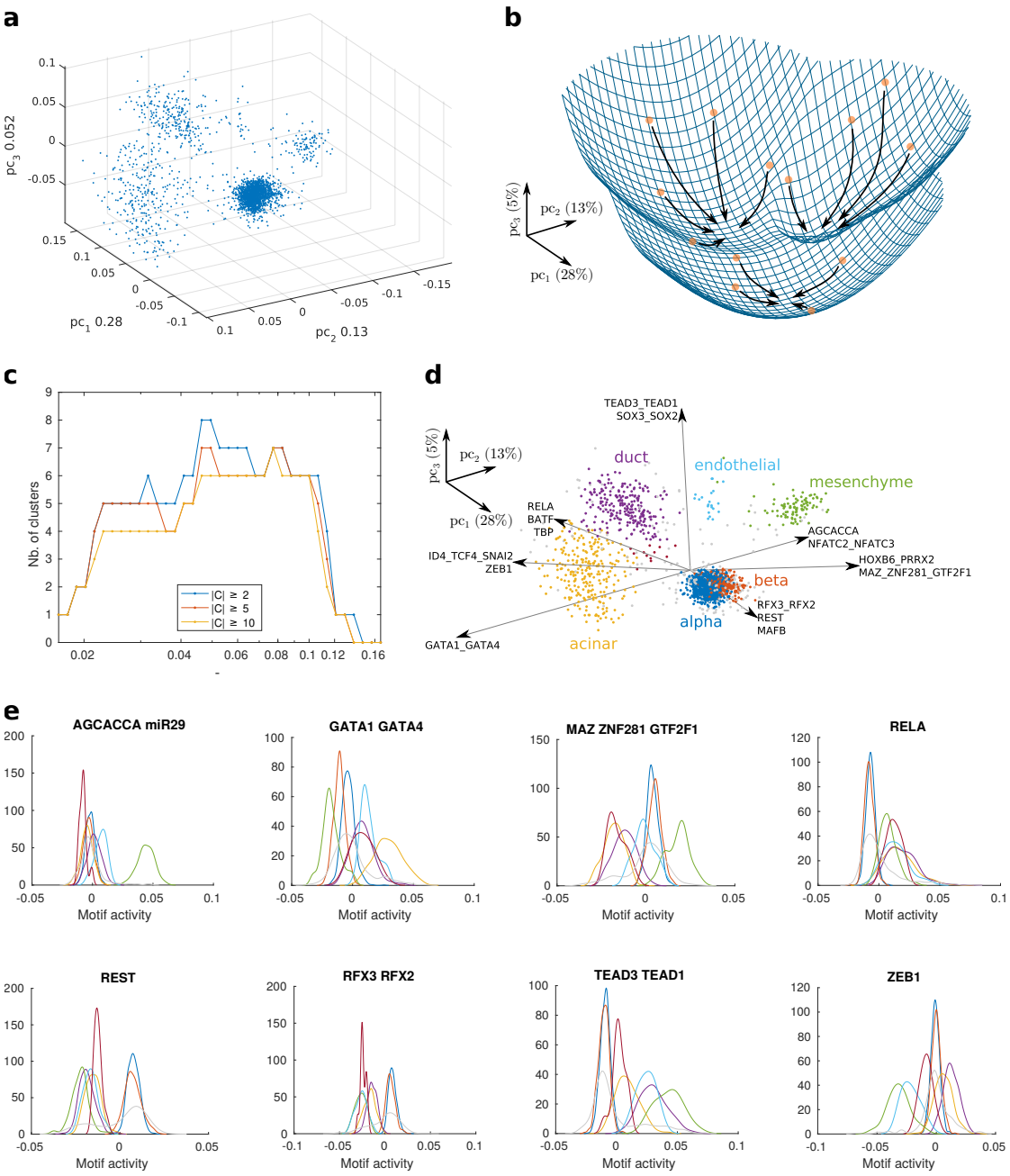


Figure 5.3: Single-cell and epigenetic landscape analysis of human pancreatic cells. **a**: 2'298 human pancreatic cells projected on the first 3 principal components. Percentage of variance captured by each component indicated on the axis. **b**: Epigenetic landscape inferred from the cells on panel a, projected on the first two principal components of panel a. **c**: Number of cluster as a function of b for clusters of at least 2 cells (blue), 5 cells (orange) and 10 cells (yellow). **d**: 2'298 human pancreatic cells projected on the first 3 principal components (same projection as panel a) and colored by inferred cell type. The 15 gene regulators that vary the most across cell types are projected and showed as axis on the 3 first principal components. **e**: Distribution of activity of 8 regulators across the different cells types colored according to the cell types defined on panel d.

5.3.2 Neural stem cells from mouse at embryonic day 13.5

One of the essential question that scRNA-seq allows to investigate is the process of cell differentiation. As the data set of the previous section consists of mature pancreatic cells, we do not expect to observe differenciation. Hence, to look at a more complex and dynamic system, we applied the described frame work (5.2) on a scRNA-seq data set of mouse neural stem cells from the forebrain cortex sampled at day 13.5, during neurogenesis. The neocortex of mouse and human is developed during neurogenesis in successive layers of distinct types of neurons, and those neurons come from the asymmetric division of the neural stem cell. Alternatively, neural stem cell undergo symmetric division giving rise to other neural stem cell [Lee et al., 2014, Mukhtar and Taylor, 2018, Paridaen and Huttner, 2014, Götz and Huttner, 2005, Taverna et al., 2014].

We applied the framework described in the previous section (see section 5.2) and inferred the landscape from those cells. Figure 5.4.a shows the 5'793 cells projected on the 3 first pc, together capturing 39% of the total variance and Figure 5.4.b show the inferred landscape projected on the first 2 principal components defined by the the cells. Applying the exact same procedure as on the previous section (section 5.3.1), we found 4 local minima on the landscape. However, in we see a distinct behaviour. Indeed, if the minima were clearly separated in section 5.3.1, we now observe a valley connecting different minima rather that clearly separated minima. As we would expect, the sampled neuronal stem cells are undergoing a differentiation process such that we must observe the fluctuations associated with the process. To observe this valley connecting the different minima, we computed the minimal energy path between every pair of those 4 minima. The minimal energy path is defined as the path that connects two points of the landscape such that the average energy along the path is minimized. To find the 6 different possible paths connecting each pair of the 4 minima, we used the zero temperature string method [E et al., 2005, Ren et al., 2005]. The zero-temperature string method is a recursive algorithm. Starting from an initial line between 2 points A and B, the initial string is defined as a group of N equidistant "beads" with the first and the last bead corresponding to A and B. As initial string, we put all N beads along the straight line joining A and B. We then compute the first derivative of the landscape at the position of each N beads and move each bead in a small step dx (parameter) following the gradient. We then do a linear interpolations along each successive N points and replace each bead but the first and the last along the interpolation lines such that the beads are equidistant. Surprisingly the 6 different possible path joining all pairs of minima all ended on a single minimal energy path connecting all 4 minima as showed as a red path on figure 5.4.b. This confirms the observation that the single cells are arranged along one main valley.

The inferred minimal energy path is defined in the whole space of regulator activity and we can ask which regulators are most variance along the path. Figure

5.4 shows the activity of the 16 most variable regulators, along the minimal energy path. We can see different groups of regulators with similar profiles. Those regulators have predicted target genes on which we performed a gene ontology analysis [Ashburner et al., 2000, Carbon et al., 2021] to ask what process are being regulated. As showed on figure 5.4.c, we identified 3 groups of regulators with similar profiles. The first group consist of Hoxb7, Sox2, Sox3, Sox10, Sox6, Sox9, Ezh2/Atf2/Ikzf1, miR-30 (GUAAACA), Pou5f1, Neurod1 and Sp1 and their target genes are related to various biological process related to neuronal differentiation. The second group consists of E2f7, E2f6, E2f2, E2f5, E2f1, Max and Mycn, whose target are related to DNA replication. The third group consist of Max, Mycn, Gcm2, Ybx1, Nfya, Nfyb, Nfyc and Cebpz, whose whose target genes are related to and cell mitosis.

This indicate that on the first half of the valley, cells are between mitosis state and S phase state, indicating cells going through cell cycle, whereas cells are going towards a state of neuronal differentiation along the rest of the valley. Those observations are in good agreement with current knowledge about the studied system where neural stem cells undergo alternatively a symmetric division to increase to neural stem cells population or an asymmetric division giving rise to basal progenitors that start differentiating towards neurons [Mukhtar and Taylor, 2018].

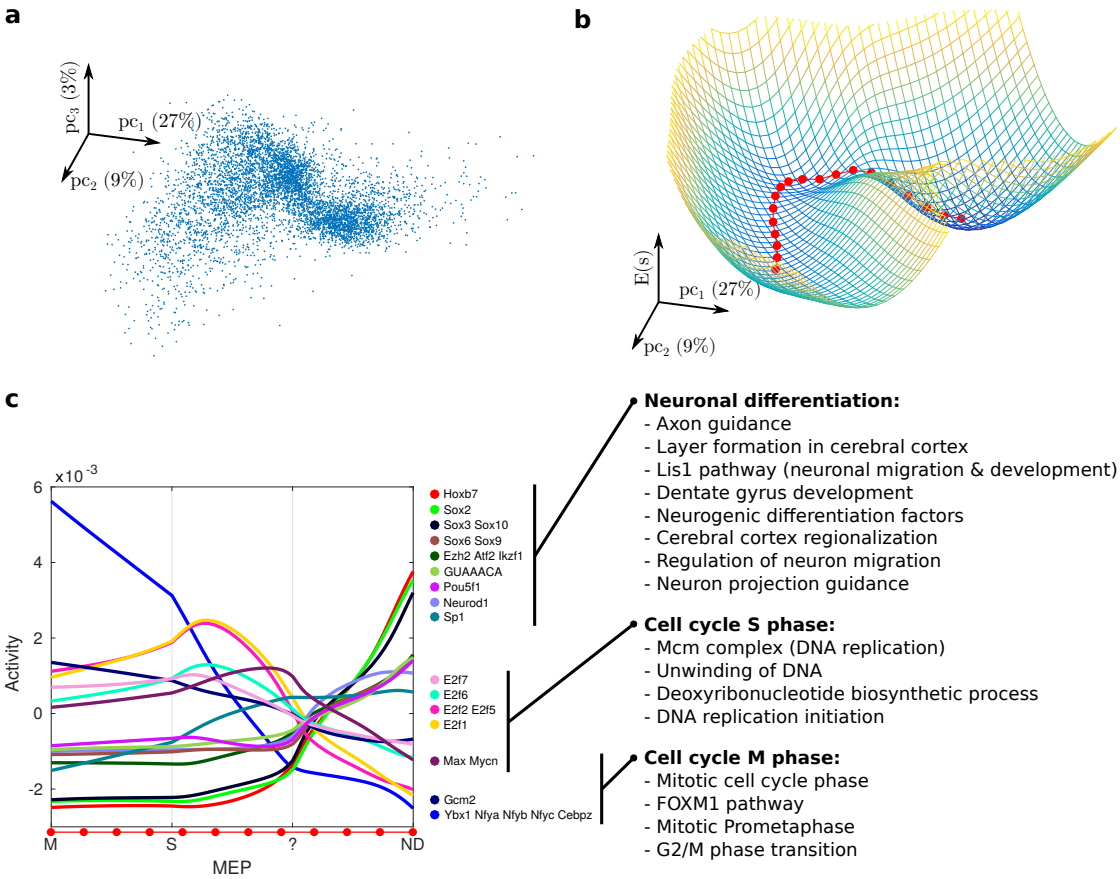


Figure 5.4: Single-cell and epigenetic landscape analysis of mouse neural stem cells. **a**: 5'793 mouse neural stem cells projected on the first 3 principal components. Percentage of variance captured by each component indicated on the axis. **b**: Epigenetic landscape inferred from the cells on panel a, projected on the first two principal components of panel a. Minimal energy path shown in red. **c**: Activity profile of the 16 gene regulators that are most variable along the minimal energy path showed on panel b. For each group of regulators that have similar profiles, we performed a gene ontology analysis [Ashburner et al., 2000, Carbon et al., 2021] on their joined predicted target genes. The biological processes associated with each group of regulators are showed.

5.4 Discussion

We presented here a procedure that uses the previously published methods Sanity [Breda et al., 2021] and ISMARA [Balwierz et al., 2014] to reconstruct an energy landscape from the density of cells in regulatory activity space. We illustrated two applications of this landscape; finding cell types as minima of the landscape and identify the regulators that distinguish those cell types, and identify developmental path with associated gene regulators. The aim wasn't here to discover new biology

about the studied systems but rather to use consensual biological knowledge to validate our method as a proof of principle.

The tasks of finding cell types and developmental trajectories in scRNA-seq data have been extensively explored, however our approach to those questions using a regulatory landscape is, to our knowledge, novel.

Among the multitude of methods currently used for the clustering of scRNA-seq we can identify a few general strategies that are based on some distance measure between each pair of cells, the arrangement of cells on a hierarchical tree, dimensional reduction, a graph representation of all cells considering a defined number of nearest neighbor cells, and often a combination of those strategies [Kiselev et al., 2017]. Bench-markings have not yet been able to point out one strategy that clearly prevails, but is it rather observed that the best performing methods are depending on the dataset [Freytag et al., 2018, Menon, 2018, Mereu et al., 2018]. This could be due to experimental, biological or technical differences between data sets used for bench-marking, and likely the lack of a clear definition of what a cell type is when constructing gold-standard data sets that are essential for a meaningful bench-marking.

Comparing methods that infer developmental trajectories is more challenging than cell clustering because we do not process ground truth data that could represent a gold-standard of developmental strategies. Each methods has to be evaluated from the biological validity of its prediction. In this task as well, we observed that our approach is novel. The reconstruction of a one-dimensional path of cellular development, often referred to as *pseudo-time*, generally consist in a dimensional reduction step followed by the trajectory modeling that infer either a linear or a branched path [Cannoodt et al., 2016, Kester and van Oudenaarden, 2018]. The most popular strategy of trajectory modeling represents all cells in a graph and find the longest path in a minimal spanning tree [Trapnell et al., 2014] or the shortest walk from a defined starting cell to every other [Matsumoto and Kiryu, 2016, Bendall et al., 2014, Setty et al., 2016, Welch et al., 2016]. Alternatively, analogous graph based methods use cell clusters rather the unique cells for constructing the graph [Shin et al., 2015, Chen et al., 2016, Ji and Ji, 2016, Grün et al., 2016]. Additional strategies include using the information of sampling time in subsequent samplings [Eugenio et al., 2014], using t-SNE transformation to fit *principal curves* [Eugenio et al., 2014], or estimating the transcription time derivative using the spliced to unspliced ratio of mRNA molecules to estimate the orientation of the trajectory in the near future in a framework known as *RNA velocity* [La Manno et al., 2018, Svensson and Pachter, 2018, Bergen et al., 2019].

One crucial point of our proposed framework consisting in inferring cell types and trajectory from a energy landscape inspired by Waddington's ideas, it that the use ISMARA to perform dimensional reduction, which produces local minima and trajectories in the space of gene regulators. This has the invaluable advantage of increasing the results interpretability. In fact, as illustrated in this work, the

gene regulators that define minima of the landscape as cell types or valleys as developmental trajectories can be interpreted as the key factor that constrain cells into a stable state or that are responsible for the changes in gene expression associated with a developmental process. Another key difference and advantage of the energy landscape as defined here is that it is twice differentiable anywhere in the space of regulatory motif activity. This allows to easily infer regulatory interaction from second derivative in a minima or along a trajectory, and to find saddle point as metastable states along developmental trajectories. Indeed, saddle points between two cell types or along a trajectory theoretically represent energy barrier that have to be passed in order to reprogram a cell from one type into another. In other words, it would hint on the key transcription factors that whose expression would need to be perturbed in reprogramming protocols [Aydin and Mazzoni, 2019, Wang et al., 2021]. Of course, this is purely theoretical and several investigations and validations are needed to support this statement.

The method used to perform the normalization of the scRNA-seq data provides error bars on the estimated gene expression per cell, and it has been shown to add valuable information for subsequent analysis [Breda et al., 2021]. Therefor, we aim in a future work to incorporate those error bars into the ISMARA model used to infer regulatory activity in single-cells and in the inference of the energy landscape. We believe this could bring a valuable improvement to the work presented here.

Chapter 6

Discussion

The general aim that links all parts of the work presented here is to elucidate the gene regulatory programs that drive cell differentiation. These are responsible for the astonishing complexity of multi-cellular organisms and also underlie complex human diseases. The first two sections of this thesis dealt with the prediction of the strength of interaction between a regulator and its target. I have focused on miRNA-mRNA interactions, as miRNAs are important post-transcriptional regulators of mRNA expression levels. Thus, accurate prediction of their targets is necessary for the design of further models that aim to explain gene expression levels in terms of the activity of their potential regulators. Ultimately, this type of model provides information about the role played by various regulators across conditions, time points or cells types [Suzuki et al., 2009, Balwierz et al., 2014]. This modelling approach takes advantage of the fact that the transcriptome-wide gene expression can be much more efficiently measured, in terms of time consumption and cost, with today's technology than the activity of regulators. The latter is technically more challenging to measure, and only for one regulator at the time. Such a model of gene regulation, called *motif activity response analysis* (in short, MARA) has been developed about a decade ago by my PhD advisers, and was used to analyse microarray and bulk RNA-seq data, to uncover TF and miRNA regulators that are active in various conditions [Balwierz et al., 2014, Suzuki et al., 2009]. An important ingredient in this analysis are prediction of regulator binding sites (TFs and miRNAs). Using known sequence specificities of TFs, whole genome sequences, and measures of conservation of potential binding site across closely related species, predictions of TF binding sites in windows of 500 base pairs upstream and downstream of the transcription start site of every annotated gene can be obtained [Arnold et al., 2012]. With a similar approach, but starting from predictions of miRNA-target interaction energies using the models that I worked on, miRNA binding sites can also be predicted genome-wide [Gumienny and Zavolan, 2015]. Assuming that the effect of every regulator on the expression of its targets is proportional to the regulator's activity, MARA infers the most likely activity of each regulator given the measured gene expression patterns across multiple samples. At a first look, the data from an scRNA-seq experiment seems similar to data from bulk RNA-seq experiments. However, treating scRNA-seq as a large collection of bulk RNA-seq data sets and applying MARA, one quickly realises that there is a major difference which hampers this analysis. This difference is in the structure of the noise, as explained in the introduction and the third part of this present work. In fact, if the noise in log-expression displayed by microarray or bulk RNA-seq data can be assumed homogeneous across genes and is dampened by the aggregation of thousands of cells, the noise in log-expression of raw scRNA-seq data depends systematically on the expression level, the most variable genes being those with lowest expression. Thus, when a model like MARA attempts to explain changes in gene expression between cells as a linear combination of changes of regulators activity, what the model is faced with is almost entirely

random fluctuations, either due to the stochastic nature of gene expression or to the experimental sampling noise. Thus, it becomes clear that reducing this noise is paramount to a meaningful analysis of regulation in single cells. It is for this precise reason that we developed Sanity, as extensively described in the third part of this thesis.

Since its initial release, the MARA model has brought valuable and validated insights about gene regulation in various systems [Gruber et al., 2014, Grunin et al., 2016, Yan et al., 2016, Dimitrova et al., 2017, Yeung et al., 2018, Tauran et al., 2019, Danoy et al., 2019]. Based on its insightful predictions at the population level, we are hopeful that applying the MARA model to single cell gene expression will help elucidate cell-specific regulatory states, and reveal regulatory interactions that drive cellular level heterogeneity, cell differentiation and development.

Applying MARA on a single scRNA-seq data set reveals the regulatory state of the cells captured in that experiment. This still does not yet reveal the dynamics of gene expression along differentiation paths. However, assuming that the cells that are captured from a complex tissue in a given experiment reflect an on-going differentiation dynamics that takes place in that tissue, and that the sampling done within the experiment is uniform, i.e. the probability of a given cell state/type to be captured is proportional to the time spent by cells in that specific state, the obtained cell states provide a fine grain description of the studied system. This should reflect the regulatory mechanisms that distinguish closely related cells undergoing a common dynamic process, being for instance, cellular differentiation of a population, reaction to a perturbation, the cell cycle, or even the stochastic fluctuations of gene expression states around a stable state. This view of a scRNA-seq sample is reminiscent of a concept proposed roughly half a century ago by Conrad Hal Waddington, namely that of an epigenetic landscape, analogous to a potential energy landscape. The epigenetic landscape is viewed as resulting from epigenetic interactions that are encoded in GRNs, constrain the regions of gene expression space that are accessible to cells and drive them through those regions towards stable states that correspond to stable cell states/types. Undoubtedly, at the time when Waddington postulated the existence of such epigenetic regulation [Waddington, 1957], none of the current knowledge about gene regulation was established, the DNA structure had just been discovered [Watson and Crick, 1953], the primary structure of proteins had just been hypothesised to be determined by the linear structure of DNA [Dounce, 1952, Gamow, 1954], the neoclassical concept of a gene was about to reach its peak with the theory of one gene controlling the synthesis of one messenger RNA, consecutively controlling the synthesis of one protein [Portin, 2002, Portin and Wilkins, 2017]. It was decades before concepts such as gene duplication, alternative splicing, overlapping genes, promoter architecture, alternative polyadenylation, and enhancers were discovered [Portin, 2002, Portin and Wilkins, 2017]. However Waddington hypothesised a "complex system of interactions underlying the epigenetic landscape" controlled by the "chemical tendencies which

genes produce" [Waddington, 1957]. This is a very general view, solely assuming that the state of a cell changes in a continuous manner and that the likelihood of a cell changing its state depends on a dynamic landscape that depends on other genes. This is also a radically different view than the previous of the epigenetic landscape as a rail structure, because it leaves room for stochasticity in the process as well as for processes such as reprogramming, which depend on an intrinsic reversibility of differentiation. Various concepts related to Waddington's epigenetic landscape have acquired specific embodiments during decades of molecular biology research. For example, gene expression regulatory process have been uncovered, cell differentiation pathways have been mapped, the stochasticity in gene expression and regulation have been demonstrated. Many aspects are still unclear, but state of the art methods in transcriptomics and epigenomics at the single cell resolution have brought us in a position where measurements at single cell resolution can reveal cellular states along a structure such as imagined by Waddington. Developing tools for this task such as the scRNA-seq MARA model is therefore both extremely valuable and very exciting.

While stochasticity was long viewed as a nuisance, more recent work started to reveal other interesting aspects. For example, work from my group has shown that the coupling between a target gene and a noisy regulator could be viewed as a rudimentary form of gene regulation [Wolf et al., 2015]. On the other hand, in my work I have found that the variability in expression between cells can also reveal important information about their underlying regulatory state. Gene expression being inherently stochastic [Mcadams and Arkin, 1997, Elowitz et al., 2002, Paulsson, 2005, Raj and van Oudenaarden, 2008], we expect gene regulatory interactions to define a probability density of cell states. By inverting this map, in the work on single cell MARA we aim to learn about the GRN-dependent landscape structure by taking advantage of the distribution of cells over gene expression states. That is, GRNs engender a function defined in gene regulatory space analogous to a potential energy which constrains the stochastic fluctuations as well as the continuous changes in the gene expression state of cells. Such a system, involving a large number of particles with high degree of freedom, with an inherently stochastic nature and under the influence of an energy potential, have been deeply described by statistical physics, and from this premise we conclude that the distribution of cells in gene expression space subject to regulatory interactions, will follow the maximum entropy distribution given only the energy function generated by GRNs, known as the Boltzmann distribution. After a rigorous inference of the state of gene expression of cells from a scRNA-seq experiment using the method Sanity presented in this work, the MARA algorithm estimate a posterior probability distribution for each cell of the measured population defined in the regulatory space of transcription factors and miRNA regulatory activity, in the form of a multivariate Gaussian distribution. Collectively considered, and assuming a uniform cell capture probability, the cells posterior probabilities create a proba-

bility distribution of cell state specific to the sampled cell population, and the hypothesis of this distribution being the maximum entropy distribution given an energy function allows us to infer this energy function. Going back to Waddington's analogy of marbles rolling down a landscape, this framework comes down to reconstructing the shape of a landscape given only the instantaneous positions of thousands of marbles rolling down the landscape at given and predefined time points. As the epigenetic potential comes from the joint distribution of thousands of cell posterior probabilities, we obtain a function defined in the entire regulatory space and differentiable in that whole space. Different applications emerge from the constructed function, tackling relevant questions of single-cell biology.

The epigenetic surface contains an intuitive definition of cell types, as the differentiability of the surface allows to computationally find local minimal. Every captured cell has an estimated position on the inferred epigenetic surface and a gradient descent algorithm assigns a unique local minimum to each cell such that every minimum defines a distinct cell type and every cell has its type assigned by the minimum it falls in. Since the early developments of scRNA-seq, the identification of cell subpopulations has probably been one of the main applications, which brought valuable insights about the cellular heterogeneity of tissues and organs [Jaitin et al., 2014, Patel et al., 2014, Zeisel et al., 2015, Baron et al., 2016, La Manno et al., 2016, Chen et al., 2017, Schaum et al., 2018]. Today there are more than a hundred published tools designed for this specific task [Zappia et al., 2018]. They cluster into a few distinct strategies, which are often combined [Kiselev et al., 2019]. Multiple methods are based on a distance measure between pairs of cells, like the popular k-means method that starts from a random initial clustering and converges to a local optimum satisfying an intra-cluster minimisation of the variance [Lloyd, 1982]. Comparably popular, hierarchical clustering methods create and entire linkage tree of the cells, with the advantage of producing different degrees of granularity and a measure of distance between clusters [Ward, 1963]. Distance-based methods are often used after dimensional reduction, which typically projects the data on the first k principal components, or after feature selection, typically filtering out genes below a predefined variance threshold function on the gene mean expression. The latter approach aims to overcome the issue known as the "curse of dimensionality" [Bellman, 1957], which denotes a concentration of distances around a common value as the dimensionality increases. This is due to the sparsity of the data points in a volume rapidly increasing with the dimension. Graph-based strategies for clustering have became increasingly popular [Kiselev et al., 2019]. The approach is to create a graph with nodes as cells and edges as a predefined number of nearest neighbours of each cell, a cluster being identified as a community of connected cells. Benchmarking studies do not uncover a globally optimal method [Freytag et al., 2018, Menon, 2018, Mereu et al., 2018], as the rankings differ across tested datasets. Thus, different methods seem to be used depending on various parameters of the sequencing experiment such as the depth

of the sequencing [Menon, 2018]. A seemingly advantageous aspect of defining cell types as minima of an epigenetic landscape, is that the shape of the surface itself in the neighbourhood of each minimum contains biological information about the regulators that are responsible for stabilising the cell state/type and, conversely, about the regulatory interactions that need to be altered to change one cell state into another.

The clustering approach in scRNA-seq data assumes cell types as discrete regions of gene expression space and implicitly neglects dynamic processes that cells undergo, such as cell cycle, cell differentiation, response to external inputs. Thus, this clustering may be appropriate when dealing with populations of cells that are fully differentiated and in 'steady state'. The identification of pathways of differentiation from scRNA-seq data has been a computational problem of great interest in recent years. Different methods have been proposed to tackle this challenge with approaches that reconstruct a one-dimensional path ordering cells according to a so-called *pseudotime*. These methods generally consist of a dimensionality reduction and a trajectory modelling component [Cannoodt et al., 2016, Kester and van Oudenaarden, 2018]. Dimensionality reduction is used to reduce the noise on the data, infer a lower dimensional manifold, select the features showing non-negligible variance, cluster together similar cells. Subsequently, differentiation trajectories between the representative types are sought, or a k-nearest neighbour graph is constructed. The graph representation is probably the most popular strategy today, and is used in various ways to infer trajectories, as the longest paths in a minimal spanning tree [Trapnell et al., 2014], or as the shortest walk in terms of added edge length from a user-defined "starting cell" to every other cell [Matsumoto and Kiryu, 2016]. The construction of a k-nearest neighbour graph is also used to similarly infer shortest path between a "starting cell" and every other cell [Bendall et al., 2014, Setty et al., 2016, Welch et al., 2016]. Graph based strategies are analogously applied on the graphs of cell clusters [Shin et al., 2015, Chen et al., 2016, Ji and Ji, 2016], which can be collected at subsequent timepoints [Eugenio et al., 2014]. On a fully connected graph constructed from clusters, each individual cell can be projected on the edges, to select edges along which the cell density is high enough to support the existence of a biologically-relevant trajectory [Grün et al., 2016]. Alternatively, it has been proposed to fit *principal curves* into lower dimensional transformation as done by t-SNE [Eugenio et al., 2014], or to estimate a time derivative of gene expression to predict the near future state of cells, and thereby an orientated trajectory as done in the *RNA velocity* framework [La Manno et al., 2018, Svensson and Pachter, 2018, Bergen et al., 2019].

The epigenetic landscape as proposed here above offers an intuitive definition of developmental trajectories as *minimum energy path* between cell states. Besides being related to Waddington's landscape, our view also relates to concepts from theoretical chemistry [Quapp and Heidrich, 1984, Sheppard et al., 2008, Vaucher and Reiher, 2018], where reactions are represented as continuous paths between

an initial and a final states along which the energy is minimised. As a matter of fact, the problem of finding the optimal path with respect to the energy in a n-dimensional energy surfaces is very general. This curve optimisation problem has been extensively studied and various algorithm have proposed [Sheppard et al., 2008]. Within the proposed epigenetic landscape framework, we assume rather intuitively, that a developmental path is a continuous region of high cell density. We assume that each cell of a population undergoing a differentiation process has equal probability to be captured in a scRNA-seq experiment. The obtained trajectory is differentiable in every dimension of the regulatory space as it is not constrained to a sequence of edges connecting cells together, but instead every point of the path takes into account the density contribution of each and every cell, with a weight exponentially decreasing with the euclidean distance as each cell has a density given by a multivariate Gaussian distribution.

An important aspect of trajectory inference is the identification of branchings that define the tree of cell fate specification. In our framework, these branchings are unambiguously determined by the minimum energy path procedure. Every point in the space of regulator activities is characterised by the weighted contribution of every cell considered in the inference of the landscape. Therefore, gene expression can also be defined everywhere in the gene regulatory space, and in particular on the minimum energy path. However, inferring developmental trajectories in the space of regulators rather than of gene expression has the considerable convenience of estimating effects that have causality embedded in them. In other words, and to reconnect with Waddington's ideas, the height of the epigenetic potential reflects physical interactions between genes and regulators; it is only relevant to assume such a potential function in a space of variables that can physical influence the gene expression state of a cell, that is the space of gene regulators activity. Hypothetically, the minimum energy connecting two distinct cell states on the epigenetic landscape, defined as local minima, gives the minimal perturbation in the level of gene regulators needed to reprogram or transdifferentiate a cell from one type to another, an application that would be highly valuable given the lack of understanding in the mechanisms underlying these processes, and the current lack of method to predict effective perturbations [Takahashi and Yamanaka, 2016].

Differences in the density of cells in the gene regulatory space can reflect differences in the relative abundance of these states, and thus, states that are very transient will be poorly represented in the data. Recent developments in scRNA-seq analysis have brought an elegant solution to the issues of sampling transient states and inferring directionality of single cell trajectories. Specifically, in RNA-seq experiments, a considerable number of captured mRNA are still unspliced, between 15% and 25% depending on the scRNA-seq protocol [La Manno et al., 2018]. These intermediates can be captured due to the priming of poly(T) oligonucleotides to intronic regions of pre-mRNAs [La Manno et al., 2018]. Unspliced mRNAs are generated by transcription (at a certain transcription rate), then spliced into mature

mRNAs that have a certain degradation rate. Thus, assuming a constant splicing rate across gene, the measured relative abundance of spliced and unspliced molecules, allows to estimate the decay rate and solve the differential equations at steady state to obtain the instantaneous time derivative of the number of mature mRNA [La Manno et al., 2018]. Further work proposed a generalisation of this framework relaxing the steady state condition, assuming gene specific transcription, splicing and decay rates, and solving the system of differential equation with a likelihood-based dynamical model [Bergen et al., 2019].

The various methods of trajectory inference mentioned here above all rely only on scRNA-seq data and are almost all based on the assumption that captured cells come from shared lineages such that similarities between expression profiles indicates closely related states along a shared lineage. However, the gene expression based similarity alone is neglecting the general fact that all the cells come from a unique ancestor and that differentiation takes place through cell division, which should generally be considered in a descriptive model of cell differentiation if one aims to better understand the phenomenon. Different experimental approach have been developed to retrace cell lineage, that can be separated into prospective and retrospective lineage tracing. Prospective techniques use the incorporation of inheritable, measurable and distinguishable traits to a subset of cells at a given time, such that the later progenitors of each marked cell can be identified by the detection of, for instance, fluorescent markers, DNA barcodes or CRISPR-Cas9-induced small modification of DNA [Kester and van Oudenaarden, 2018]. Retrospective lineage tracing tackle a more challenging task of reconstruction the whole phylogeny of a cell population by inferring a tree from somatic mutations, copy number variation, single nucleotide variants or epigenetic marks [Kester and van Oudenaarden, 2018]. Most of those methods are informative about the genetic relationship between cells but all lack a more phenotypic characterisation that can be provided by scRNA-seq. A few methods have combined transcriptomic measurements with prospective lineage tracing, using imaging of single-molecule fluorescent *in situ* hybridization data [Hormoz et al., 2016, Kimmerling et al., 2016, Frieda et al., 2017], viral barcoding of the 3' UTR of a fluorescent protein mRNA [Yao et al., 2017, Biddy et al., 2018], or CRISPR-Cas9-induced small modifications of a fluorecent protein mRNA that can be detected in sequencing [Spanjaard et al., 2018, Raj et al., 2018, Alemany et al., 2018, Kalhor et al., 2017]. Remarkably, single-molecule fluorescent *in situ* hybridisation data also allows the spatial localisation of cells, adding a s layer of information about the spatial organisation of cells during differentiation.

More generally, the number of high-throughput experimental techniques bringing various independent measurements at the single cell resolution has exploded, providing means to uncover information about the lineage, genome sequence and methylation state, histone modification, chromatin accessibility, composition in mRNA, proteins and surface proteins, and the spatial localisation of cells in the

body [Stuart and Satija, 2019]. Some of those techniques have also been combined into one experiment, such that observations can be made simultaneously on the same cell. Most often though, the measurable variables are obtained in a destructive manner, and thereby have to be considered in distinct cell populations. However, the scale and diversity of the measurable variables describing the state of a cell leads us to believe that we are closer than ever to fully elucidate the molecular mechanisms underlying the developments of multicellular organisms, and to start building a quantitative theory of how the genetic information is interpreted to produce the astonishing complexity of a living organism. In fact, considerable efforts in the scientific community are focused towards tackling the integration of the abundance of large scaled data generated in recent years [Adey, 2019, Stuart and Satija, 2019, Welch et al., 2019], and it seems unavoidable that those efforts will have to primarily go towards the rigorous definition, characterisation, and analysis of the measured variables, before they can possibly be integrated together as different aspects of one theory.

The work achieved during the beginning of my doctoral studies and presented in the first two sections of this thesis (*Quantifying the strength of miRNA-target interactions*, and *Single cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction*), contributed to the field of miRNA-dependent regulation by improving the characterisation and the quantification of biophysical miRNA-mRNA interactions from two different view points, using measurements of the kinetics of mRNA degradation increase induced by miRNAs and the response in mRNA level to changes in miRNA concentration measured with scRNA-seq. The last part of the thesis (*Bayesian inference of gene expression levels in single cells*), presents a framework that was first designed to rigorously normalise mRNA count data from scRNA-seq experiments, as we could not find any satisfactory published solution to perform this task, which needs to be solved before applying the MARA model that has been designed in the group for bulk RNA-seq about 10 years ago [Suzuki et al., 2009, Balwierz et al., 2014]. We rapidly realised that our work could be highly relevant for any sort of analysis that needs to normalise for the Poisson distributed part of the variance as well as the fluctuations in the total number of mRNA molecules captured from a cell. As a matter of fact, our Bayesian framework solves the very general mathematical problem of estimating log-normally distributed variables presenting additional and undesirable Poisson distributed noise (typically sampling noise). Because of its generality, our work has been well received when presented in scientific conferences and since we made it available as a pre-print. As we developed it in an integrable design, our algorithm can be easily included into new workflows, extend existing ones, to improve the analysis of scRNA-seq data sets and increase the breadth and accuracy of conclusions that can be drawn from such data. The work that builds on the Sanity framework and is still unpublished, of inferring the activities of regulators of gene expression in single cells is an entirely new approach to scRNA-seq analysis, which

already generated compelling preliminary results concerning stem cell differentiation and the development of various organs such as the mouse cortex. For these reasons, we expect that this work will also be of great interest to the field, opening new avenues to engineer developmental fates.

Bibliography

[10X Genomics, 2018] 10X Genomics (2018). What fraction of mRNA transcripts are captured per cell? – 10X Genomics.

[Adamson et al., 2016] Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nunez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., and Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.

[Adey, 2019] Adey, A. C. (2019). Integration of Single-Cell Genomics Datasets. *Cell*, 177(7):1677–1679.

[Aghdassi et al., 2012] Aghdassi, A., Sendler, M., Guenther, A., Mayerle, J., Behn, C. O., Heidecke, C. D., Friess, H., Büchler, M., Evert, M., Lerch, M. M., and Weiss, F. U. (2012). Recruitment of histone deacetylases HDAC1 and HDAC2 by the transcriptional repressor ZEB1 downregulates E-cadherin expression in pancreatic cancer. *Gut*, 61(3):439–448.

[Ait-Lounis et al., 2007] Ait-Lounis, A., Baas, D., Barras, E., Benadiba, C., Charollais, A., Nlend, R. N., Liègeois, D., Meda, P., Durand, B., and Reith, W. (2007). Novel function of the ciliogenic transcription factor RFX3 in development of the endocrine pancreas. *Diabetes*, 56(4):950–959.

[Alemany et al., 2018] Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., and Van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112.

[Algül et al., 2007] Algül, H., Treiber, M., Lesina, M., Nakhai, H., Saur, D., Geisler, F., Pfeifer, A., Paxian, S., and Schmid, R. M. (2007). Pancreas-specific RelA/p65 truncation increases susceptibility of acini to inflammation-associated cell death following cerulein pancreatitis. *The Journal of clinical investigation*, 117(6):1490–1501.

[AlJanahi et al., 2018] AlJanahi, A. A., Danielsen, M., and Dunbar, C. E. (2018). An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy - Methods and Clinical Development*, 10:189–196.

[Altschuler and Wu, 2010] Altschuler, S. J. and Wu, L. F. (2010). Cellular Heterogeneity: Do Differences Make a Difference?

[Alwine et al., 1977] Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5350–4.

[Angermueller et al., 2016] Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 13(3):229–232.

[Arnold et al., 2012] Arnold, P., Erb, I., Pachkov, M., Molina, N., and Van Nimwegen, E. (2012). MotEvo: Integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, 28(4):487–494.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29.

[Aydin and Mazzoni, 2019] Aydin, B. and Mazzoni, E. O. (2019). Cell Reprogramming: The Many Roads to Success. *https://doi.org/10.1146/annurev-cellbio-100818-125127*, 35:433–452.

[Bajic et al., 2004] Bajic, V. B., Sin, L. T., Suzuki, Y., and Sugano, S. (2004). Promoter prediction analysis on the whole human genome. *Nature Biotechnology*, 22(11):1467–1473.

[Balwierz et al., 2014] Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., and Van Nimwegen, E. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, 24(5):869–884.

[Baron et al., 2016] Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human

and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*.

[Bartel, 2009] Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233.

[Bartlett and Stirling, 2003] Bartlett, J. M. S. and Stirling, D. (2003). A Short History of the Polymerase Chain Reaction. In *PCR Protocols*, pages 3–6. Humana Press, New Jersey.

[Bazzini et al., 2012] Bazzini, A. A., Lee, M. T., and Giraldez, A. J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science (New York, N.Y.)*, 336(6078):233–237.

[Beal, 2017] Beal, J. (2017). Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology*, 1(1):55–60.

[Becker-André and Hahlbrock, 1989] Becker-André, M. and Hahlbrock, K. (1989). Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcipt titration assay (PATTY). *Nucleic Acids Research*, 17(22):9437–9446.

[Bellman, 1957] Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, 1957, Princeton, New Jersey.

[Bendall et al., 2014] Bendall, S. C., Davis, K. L., Amir, E. A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'Er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725.

[Bergen et al., 2019] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2019). Generalizing RNA velocity to transient cell states through dynamical modeling. *bioRxiv*, page 820936.

[Bergen et al., 2020] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*

[Biddy et al., 2018] Biddy, B. A., Kong, W., Kamimoto, K., Guo, C., Waye, S. E., Sun, T., and Morris, S. A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature*, 564(7735):219–224.

[Bissels et al., 2009] Bissels, U., Wild, S., Tomiuk, S., Holste, A., Hafner, M., Tuschl, T., and Bosio, A. (2009). Absolute quantification of microRNAs by using a universal reference. *RNA (New York, N.Y.)*, 15(12):2375–2384.

[Blevins et al., 2015] Blevins, R., Bruno, L., Carroll, T., Elliott, J., Marcais, A., Loh, C., Hertweck, A., Krek, A., Rajewsky, N., Chen, C. Z., Fisher, A. G., and Merkenschlager, M. (2015). microRNAs regulate cell-to-cell variability of endogenous target gene expression in developing mouse thymocytes. *PLoS genetics*, 11(2):1–19.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Bornkamm et al., 2005] Bornkamm, G. W., Berens, C., Kuklik-Roos, C., Bechet, J. M., Laux, G., Bachl, J., Korndoerfer, M., Schlee, M., Hölzel, M., Malamoussi, A., Chapman, R. D., Nimmerjahn, F., Mautner, J., Hillen, W., Bujard, H., and Feuillard, J. (2005). Stringent doxycycline-dependent control of gene activities using an episomal one-vector system. *Nucleic acids research*, 33(16):1–11.

[Bosia et al., 2013] Bosia, C., Pagnani, A., and Zecchina, R. (2013). Modelling Competing Endogenous RNA Networks. *PloS one*, 8(6):e66609.

[Bosia et al., 2015] Bosia, C., Sgrò, F., Conti, L., Baldassi, C., Cavallo, F., Di Cunto, F., Turco, E., Pagnani, A., and Zecchina, R. (2015). Quantitative study of crossregulation, noise and synchronization between microRNA targets in single cells. *Genome Biology*, 18(1).

[Bosson et al., 2014] Bosson, A. D., Zamudio, J. R., and Sharp, P. A. (2014). Endogenous miRNA and Target Concentrations Determine Susceptibility to Potential ceRNA Competition. *Molecular Cell*.

[Boyle et al., 2008] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322.

[Bradley and Moxon, 2017] Bradley, T. and Moxon, S. (2017). An assessment of the next generation of animal miRNA target prediction algorithms. In *Methods in Molecular Biology*, volume 1580, pages 175–191. Humana Press Inc.

[Braun et al., 2012] Braun, J. E., Huntzinger, E., and Izaurralde, E. (2012). A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. *Cold Spring Harbor perspectives in biology*, 4(12):a012328–.

[Bray et al., 2016] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527.

[Breda et al., 2021] Breda, J., Zavolan, M., and van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 39(8):1008–1016.

[Brennecke et al., 2013] Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1098.

[Buchler and Louis, 2008] Buchler, N. E. and Louis, M. (2008). Molecular titration and ultrasensitivity in regulatory networks. *Journal of molecular biology*, 384(5):1106–19.

[Buenrostro et al., 2013] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.

[Buenrostro et al., 2015] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

[Buettner et al., 2015] Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160.

[Bulmer, 1991] Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3):897–907.

[Bulyk, 2004] Bulyk, M. L. (2004). Computational prediction of transcription-factor binding site locations.

[Bussemaker et al., 2001] Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–171.

[Cannoodt et al., 2016] Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology*, 46(11):2496–2506.

[Carbon et al., 2021] Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan,

A., Huang, X., Mushayahama, T., LaBonte, S. A., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Garapati, P., Marygold, S. J., Trovisco, V., dos Santos, G., Falls, K., Tabone, C., Zhou, P., Goodman, J. L., Strelets, V. B., Thurmond, J., Garmiri, P., Ishtiaq, R., Rodríguez-López, M., Acencio, M. L., Kuiper, M., Lægreid, A., Logie, C., Lovering, R. C., Kramarz, B., Saverimuttu, S. C., Pinheiro, S. M., Gunn, H., Su, R., Thurlow, K. E., Chibucos, M., Giglio, M., Nadendla, S., Munro, J., Jackson, R., Duesbury, M. J., Del-Toro, N., Meldal, B. H., Paneerselvam, K., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Chang, H. Y., Finn, R. D., Mitchell, A. L., Rawlings, N. D., Richardson, L., Sangrador-Vegas, A., Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D. M., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Hayles, J., Bähler, J., Bolton, E. R., de Pons, J. L., Dwinell, M. R., Hayman, G. T., Kaldunski, M. L., Kwitek, A. E., Laulederkind, S. J., Plasterer, C., Tutaj, M. A., Vedi, M., Wang, S. J., D'Eustachio, P., Matthews, L., Balhoff, J. P., Aleksander, S. A., Alexander, M. J., Cherry, J. M., Engel, S. R., Gondwe, F., Karra, K., Miyasato, S. R., Nash, R. S., Simison, M., Skrzypek, M. S., Weng, S., Wong, E. D., Feuermann, M., Gaudet, P., Morgat, A., Bakker, E., Berardini, T. Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C. N., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Blatter, M. C., Boutet, E., Bowler, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casas, C. C., Coudert, E., Denny, P., Es-Treicher, A., Famiglietti, M. L., Georghiou, G., Gos, A. N., Gruaz-Gumowski, N., Hatton-Ellis, E., Hulo, C., Ignatchenko, A., Jungo, F., Laiho, K., Le Mercier, P., Lieberherr, D., Lock, A., Lussi, Y., MacDougall, A., Ma-Grane, M., Martin, M. J., Masson, P., Natale, D. A., Hyka-Nouspikel, N., Orchard, S., Pedruzzi, I., Pourcel, L., Poux, S., Pundir, S., Rivoire, C., Speretta, E., Sundaram, S., Tyagi, N., Warner, K., Zaru, R., Wu, C. H., Diehl, A. D., Chan, J. N., Grove, C., Lee, R. Y., Muller, H. M., Raciti, D., van Auken, K., Sternberg, P. W., Berriman, M., Paulini, M., Howe, K., Gao, S., Wright, A., Stein, L., Howe, D. G., Toro, S., Westerfield, M., Jaiswal, P., Cooper, L., and Elser, J. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research*, 49(D1):D325–D334.

[Carninci et al., 2006] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635.

[Cebola et al., 2015] Cebola, I., Rodríguez-Seguí, S. A., Cho, C. H., Bessa, J., Rovira, M., Luengo, M., Chhatriwala, M., Berry, A., Ponsa-Cobas, J., Maestro, M. A., Jennings, R. E., Pasquali, L., Morán, I., Castro, N., Hanley, N. A., Gomez-Skarmeta, J. L., Vallier, L., and Ferrer, J. (2015). TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. *Nature cell biology*, 17(5):615–626.

[Cesana et al., 2011] Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369.

[Chai et al., 2014] Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48(1):55–65.

[Chandradoss et al., 2015] Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., Macrae, I. J., and Joo, C. (2015). A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, 162(1):96–107.

[Chang et al., 2004] Chang, J., Nicolas, E., Marks, D., Sander, C., Lerro, A., Buendia, M. A., Xu, C., Mason, W. S., Moloshok, T., Bort, R., Zaret, K. S., and Taylor, J. M. (2004). miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1. *RNA biology*, 1(2):106–113.

[Chen et al., 2016] Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F., and Poidinger, M. (2016). Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature Communications*, 7:1–15.

[Chen et al., 2017] Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell Reports*.

[Chi et al., 2009] Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86.

[chun Cheng and Lin, 2013] chun Cheng, E. and Lin, H. (2013). Repressing the repressor: a lincRNA as a MicroRNA sponge in embryonic stem cell self-renewal. *Developmental cell*, 25(1):1–2.

[Clapier and Cairns, 2009] Clapier, C. R. and Cairns, B. R. (2009). The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry*, 78(1):273–304.

[Clark et al., 2018] Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun*, 9(1):781.

[Coifman and Lafon, 2006] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*.

[Collins et al., 2003] Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003). A vision for the future of genomics research.

[Costa-Silva et al., 2017] Costa-Silva, J., Domingues, D., and Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, 12(12):e0190152.

[Crawford et al., 2006] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., and Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, 16(1):123–131.

[Cui and Zhao, 2012] Cui, K. and Zhao, K. (2012). Genome-Wide Approaches to Determining Nucleosome Occupancy in Metazoans Using MNase-Seq. In *Chromatin Remodeling*, chapter Genome-Wid, pages 413–419. Humana Press.

[Cusanovich et al., 2015] Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, 348(6237):910–4.

[Dahiya et al., 2008] Dahiya, N., Sherman-Baust, C. A., Wang, T. L., Davidson, B., Shih, L. M., Zhang, Y., Wood, W., Becker, K. G., and Morin, P. J. (2008). MicroRNA expression and identification of putative miRNA targets in ovarian cancer. *PloS one*, 3(6).

[Danoy et al., 2019] Danoy, M., Bernier, M. L., Kimura, K., Poulain, S., Kato, S., Mori, D., Kido, T., Plessy, C., Kusuhara, H., Miyajima, A., Sakai, Y., and Leclerc, E. (2019). Optimized protocol for the hepatic differentiation of induced pluripotent stem cells in a fluidic microenvironment. *Biotechnology and Bioengineering*, 116(7):1762–1776.

[Darnell, 2010] Darnell, R. B. (2010). HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdisciplinary Reviews: RNA*, 1(2):266–286.

[Datlinger et al., 2017] Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, 14(3):297–301.

[Davidson and Levin, 2005] Davidson, E. and Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4935.

[Davidson, 2010] Davidson, E. H. (2010). Emerging properties of animal gene regulatory networks.

[Davidson and Erwin, 2006] Davidson, E. H. and Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans.

[Davidson et al., 2019] Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J.-M. (2019). DNA loop extrusion by human cohesin. *Science (New York, N.Y.)*.

[Davis et al., 1987] Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6):987–1000.

[Denzler et al., 2014] Denzler, R., Agarwal, V., Stefano, J., Bartel, D. P., and Stoffel, M. (2014). Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. *Molecular Cell*, 54(5):766–776.

[Dimitrova et al., 2017] Dimitrova, Y., Gruber, A. J., Mittal, N., Ghosh, S., Dimitriades, B., Mathow, D., Grandy, W. A., Christofori, G., and Zavolan, M. (2017). TFAP2A is a component of the ZEB1/2 network that regulates TGFB1-induced epithelial to mesenchymal transition. *Biology Direct*, 12(1):8.

[Dixit et al., 2016] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[Dooley et al., 2016] Dooley, J., Garcia-Perez, J. E., Sreenivasan, J., Schlenner, S. M., Vangoitsenhoven, R., Papadopoulou, A. S., Tian, L., Schonefeldt, S., Serneels, L., Deroose, C., Staats, K. A., Van Der Schueren, B., De Strooper,

B., McGuinness, O. P., Mathieu, C., and Liston, A. (2016). The microRNA-29 Family Dictates the Balance Between Homeostatic and Pathological Glucose Handling in Diabetes and Obesity. *Diabetes*, 65(1):53–61.

[Dounce, 1952] Dounce, A. (1952). Duplicating mechanism for peptide chain and nucleic acid synthesis. *Enzymologia*, 15(5):251–258.

[E et al., 2005] E, W., Ren, W., and Vanden-Eijnden, E. (2005). Finite temperature string method for the study of rare events. *Journal of Physical Chemistry B*, 109(14):6688–6693.

[Ebert and Sharp, 2012] Ebert, M. S. and Sharp, P. A. (2012). Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–524.

[Eberwine et al., 1992] Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7):3010–4.

[Eichhorn et al., 2014] Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., hao Hsu, S., Ghoshal, K., Villén, J., and Bartel, D. P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Molecular cell*, 56(1):104–115.

[Elgin, 1988] Elgin, S. C. R. (1988). The Formation and Function of DNase I Hypersensitive Sites in the Process of Gene Activation. *Journal of Biological Chemistry*, 263(36):19259–19262.

[Elkayam et al., 2012] Elkayam, E., Kuhn, C. D., Tocilj, A., Haase, A. D., Greene, E. M., Hannon, G. J., and Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. *Cell*, 150(1):100–110.

[Elowitz et al., 2002] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

[Engvall and Perlmann, 1972] Engvall, E. and Perlmann, P. (1972). Enzyme-linked immunosorbent assay, Elisa. 3. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes. *Journal of immunology (Baltimore, Md. : 1950)*, 109(1):129–35.

[Eraslan et al., 2019] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390.

[Erhard et al., 2014] Erhard, F., Haas, J., Lieber, D., Malterer, G., Jaskiewicz, L., Zavolan, M., Dölken, L., and Zimmer, R. (2014). Widespread context dependency of microRNA-mediated regulation. *Genome research*, 24(6):906–919.

[Eugenio et al., 2014] Eugenio, M., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G. C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 111(52):E5643–E5650.

[Fan et al., 2016] Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J. B., Zhang, K., Chun, J., and Kharchenko, P. V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3):241–244.

[Farh et al., 2005] Farh, K. K. H., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B., and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science (New York, N.Y.)*, 310(5755):1817–1821.

[Figliuzzi et al., 2013] Figliuzzi, M., Marinari, E., and De Martino, A. (2013). MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory. *Biophysical journal*, 104(5):1203–13.

[Forrest et al., 2014] Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Lassmann, T., Itoh, M., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70.

[Frankel et al., 2011] Frankel, L. B., Wen, J., Lees, M., Høyer-Hansen, M., Farkas, T., Krogh, A., Jäättelä, M., and Lund, A. H. (2011). microRNA-101 is a potent inhibitor of autophagy. *The EMBO journal*, 30(22):4628–4641.

[Freeman et al., 1999] Freeman, W. M., Walker, S. J., and Vrana, K. E. (1999). Quantitative RT-PCR: Pitfalls and Potential. *BioTechniques*, 26(1):112–125.

[Frei et al., 2016] Frei, A. P., Bava, F. A., Zunder, E. R., Hsieh, E. W., Chen, S. Y., Nolan, G. P., and Gherardini, P. F. (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods*, 13(3):269–275.

[Freytag et al., 2018] Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7:1297.

[Frieda et al., 2017] Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K. H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B., and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111.

[Friedman et al., 2006] Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Physical Review Letters*, 97(16).

[Gaidatzis et al., 2007] Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8(1):69.

[Gamow, 1954] Gamow, G. (1954). Possible relation between deoxyribonucleic acid and protein structures.

[Gao, 2018] Gao, S. (2018). Data Analysis in Single-Cell Transcriptome Sequencing. *Methods in molecular biology (Clifton, N.J.)*, 1754:311–326.

[Garcia et al., 2011] Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., and Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature structural & molecular biology*, 18(10):1139–46.

[Gennarino et al., 2009] Gennarino, V. A., Sardiello, M., Avellino, R., Meola, N., Maselli, V., Anand, S., Cutillo, L., Ballabio, A., and Banfi, S. (2009). MicroRNA target prediction by expression analysis of host genes. *Genome research*, 19(3):481–90.

[Genomics, 2020] Genomics, X. (2020). Cell Ranger DNA. https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/what-is-cell-ranger-dna.

[Gershoni and Palade, 1983] Gershoni, J. M. and Palade, G. E. (1983). Protein Blotting: Principles and Applications. *Analytical Biochemistry*, 131:1–15.

[Ghosh et al., 2015] Ghosh, S., Bose, M., Ray, A., and Bhattacharyya, S. N. (2015). Polysome arrest restricts miRNA turnover by preventing exosomal export of miRNA in growth-retarded mammalian cells. *Molecular biology of the cell*, 26(6):1072–1083.

[Giresi et al., 2007] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877–85.

[Gomez et al., 2013] Gomez, D., Shankman, L. S., Nguyen, A. T., and Owens, G. K. (2013). Detection of histone modifications at specific gene loci in single cells in histological sections. *Nature Methods*, 10(2):171–177.

[Götz and Huttner, 2005] Götz, M. and Huttner, W. B. (2005). The cell biology of neurogenesis. *Nature reviews. Molecular cell biology*, 6(10):777–788.

[Goutsias and Lee, 2007] Goutsias, J. and Lee, N. (2007). Computational and Experimental Approaches for Modeling Gene Regulatory Networks. *Current Pharmaceutical Design*, 13(14):1415–1436.

[Grosswendt et al., 2014] Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Molecular cell*, 54(6):1042–1054.

[Gruber et al., 2014] Gruber, A. J., Grandy, W. A., Balwierz, P. J., Dimitrova, Y. A., Pachkov, M., Ciaudo, C., van Nimwegen, E., and Zavolan, M. (2014). Embryonic stem cell-specific microRNAs contribute to pluripotency by inhibiting regulators of multiple differentiation pathways. *Nucleic Acids Research*, 42(14):9313–9326.

[Gruber and Zavolan, 2013] Gruber, A. J. and Zavolan, M. (2013). Modulation of epigenetic regulators and cell fate decisions by miRNAs. *Epigenomics*, 5(6):671–83.

[Gruber and Zavolan, 2019] Gruber, A. J. and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, page 1.

[Grün et al., 2014] Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640.

[Grün et al., 2015] Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and Van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.

[Grün et al., 2016] Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., de Koning, E. J., and van Oudenaarden, A. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19(2):266–277.

[Grün et al., 2005] Grün, D., Wang, Y. L., Langenberger, D., Gunsalus, K. C., and Rajewsky, N. (2005). microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS computational biology*, 1(1):0051–0066.

[Grunin et al., 2016] Grunin, M., Shira-Hagbi-Levi, Rinsky, B., Smith, Y., and Chowers, I. (2016). Transcriptome analysis on monocytes from patients with neovascular age-related macular degeneration. *Scientific Reports*, 6.

[Guimaraes et al., 2014] Guimaraes, J. C., Rocha, M., and Arkin, A. P. (2014). Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Research*, 42(8):4791–4799.

[Gumienny and Zavolan, 2015] Gumienny, R. and Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Research*, 43(3):1380–1391.

[Guo et al., 2010] Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840.

[Ha et al., 1996] Ha, I., Wightman, B., and Ruvkun, G. (1996). A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes & development*, 10(23):3041–3050.

[Ha and Kim, 2014] Ha, M. and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology*, 15(8):509–524.

[Hafemeister and Satija, 2019] Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, 20(1):296.

[Hafner et al., 2010] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141.

[Haghverdi et al., 2016] Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848.

[Hannenhalli, 2008] Hannenhalli, S. (2008). Eukaryotic transcription factor binding sites - Modeling and integrative search methods.

[Hansen et al., 2013] Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–388.

[Hashimshony et al., 2012] Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, 2(3):666–673.

[Hausser et al., 2009] Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M. (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome research*, 19(11):2009–2020.

[Hausser et al., 2013] Hausser, J., Syed, A. P., Selevsek, N., Van Nimwegen, E., Jaskiewicz, L., Aebersold, R., and Zavolan, M. (2013). Timescales and bottlenecks in miRNA-dependent gene regulation. *Molecular systems biology*, 9.

[Hausser and Zavolan, 2014] Hausser, J. and Zavolan, M. (2014). Identification and consequences of miRNA-target interactions - beyond repression of gene expression. *Nature reviews. Genetics*, 15(9):599–612.

[Heinz et al., 2010] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589.

[Helwak et al., 2013] Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–65.

[Hofacker et al., 1994] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie Chemical Monthly*, 125(2):167–188.

[Hormoz et al., 2016] Hormoz, S., Singer, Z. S., Linton, J. M., Antebi, Y. E., Shraiman, B. I., and Elowitz, M. B. (2016). Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Systems*, 3(5):419–433.e8.

[Hornstein and Shomron, 2006] Hornstein, E. and Shomron, N. (2006). Canalization of development by microRNAs. *Nature genetics*, 38 Suppl(6S):S20.

[Hoyle et al., 2002] Hoyle, D. C., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–584.

[Huang et al., 2018] Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542.

[Hudson et al., 2013] Hudson, R. S., Yi, M., Esposito, D., Glynn, S. A., Starks, A. M., Yang, Y., Schetter, A. J., Watkins, S. K., Hurwitz, A. A., Dorsey, T. H.,

Stephens, R. M., Croce, C. M., and Ambs, S. (2013). MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer. *Oncogene*, 32(35):4139–4147.

[Huntzinger and Izaurralde, 2011] Huntzinger, E. and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature reviews. Genetics*, 12(2):99–110.

[Huynh-Thu and Sanguinetti, 2019] Huynh-Thu, V. A. and Sanguinetti, G. (2019). Gene Regulatory Network Inference: An Introductory Survey. In *Gene Regulatory Networks*, chapter Gene Regul, pages 1–23. Humana Press, New York, NY.

[Islam et al., 2014] Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166.

[Jackson et al., 2006] Jackson, A. L., Burchard, J., Schelter, J., Chau, B. N., Cleary, M., Lim, L., and Linsley, P. S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA (New York, N.Y.)*, 12(7):1179–1187.

[Jacob and Monod, 1961] Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356.

[Jaitin et al., 2014] Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

[Jaitin et al., 2016] Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7):1883–1896.

[Jaskiewicz et al., 2012] Jaskiewicz, L., Bilen, B., Hausser, J., and Zavolan, M. (2012). Argonaute CLIP–a method to identify in vivo targets of miRNAs. *Methods (San Diego, Calif.)*, 58(2):106–112.

[Jayaram et al., 2016] Jayaram, N., Usvyat, D., and R. Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*.

[Jaynes, 2003] Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

[Jensen and Darnell, 2008] Jensen, K. B. and Darnell, R. B. (2008). CLIP: Crosslinking and ImmunoPrecipitation of In Vivo RNA Targets of RNA-Binding Proteins. In *RNA-Protein Interaction Protocols*, chapter CLIP: Cros, pages 85–98. Humana Press.

[Ji and Ji, 2016] Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117.

[Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502.

[Jolliffe, 2005] Jolliffe, I. (2005). Principal Component Analysis. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, Chichester, UK.

[Kalhor et al., 2018] Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., and Church, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science*, 361(6405).

[Kalhor et al., 2017] Kalhor, R., Mali, P., and Church, G. M. (2017). Rapidly evolving homing CRISPR barcodes. *Nature Methods*, 14(2):195–200.

[Kanellopoulou et al., 2005] Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D. M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes & development*, 19(4):489–501.

[Karreth et al., 2015] Karreth, F. A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., Sjoberg, M., Keane, T. M., Verma, A., Ala, U., Tay, Y., Wu, D., Seitzer, N., Velasco-Herrera, M. D. C., Bothmer, A., Fung, J., Langellotto, F., Rodig, S. J., Elemento, O., Shipp, M. A., Adams, D. J., Chiarle, R., and Pandolfi, P. P. (2015). The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell*, 161(2):319–332.

[Katahira et al., 2015] Katahira, J., Katahira, and Jun (2015). Nuclear Export of Messenger RNA. *Genes*, 6(2):163–184.

[Kester and van Oudenaarden, 2018] Kester, L. and van Oudenaarden, A. (2018). Single-Cell Transcriptomics Meets Lineage Tracing.

[Khan et al., 2009] Khan, A. A., Betel, D., Miller, M. L., Sander, C., Leslie, C. S., and Marks, D. S. (2009). Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nature biotechnology*, 27(6):549–555.

[Khorshid et al., 2013] Khorshid, M., Hausser, J., Zavolan, M., and van Nimwegen, E. (2013). A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10(3):253–5.

[Kim et al., 2010] Kim, Y. K., Heo, I., and Kim, V. N. (2010). Modifications of small RNAs and their associated proteins. *Cell*, 143(5):703–709.

[Kimmerling et al., 2016] Kimmerling, R. J., Lee Szeto, G., Li, J. W., Genshaft, A. S., Kazer, S. W., Payer, K. R., De Riba Borrajo, J., Blainey, P. C., Irvine, D. J., Shalek, A. K., and Manalis, S. R. (2016). A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nature Communications*, 7.

[Kimura et al., 2006] Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J. I., Sekine, M., Tsuritani, K., Wakaguri, H., Ishii, S., Sugiyama, T., Saito, K., Isono, Y., Irie, R., Kushida, N., Yoneyama, T., Otsuka, R., Kanda, K., Yokoi, T., Kondo, H., Wagatsuma, M., Murakawa, K., Ishida, S., Ishibashi, T., Takahashi-Fujii, A., Tanase, T., Nagai, K., Kikuchi, H., Nakai, K., Isogai, T., and Sugano, S. (2006). Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Research*, 16(1):55–65.

[Kiselev et al., 2019] Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282.

[Kiselev et al., 2017] Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486.

[Kishore et al., 2011] Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7):559–567.

[Klein et al., 2015] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

[Kornblihtt, 2005] Kornblihtt, A. R. (2005). Promoter usage and alternative splicing.

[Kramer, 1991] Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

[Kulaeva et al., 2012] Kulaeva, O. I., Nizovtseva, E. V., Polikanov, Y. S., Ulianov, S. V., and Studitsky, V. M. (2012). Distant Activation of Transcription: Mechanisms of Enhancer Action. *Molecular and Cellular Biology*, 32(24):4892–4897.

[Kulakovskiy et al., 2013] Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013). HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research*, 41(D1).

[Kulessa et al., 1995] Kulessa, H., Frampton, J., and Graf, T. (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes & development*, 9(10):1250–62.

[Kundaje et al., 2016] Kundaje, A., Boley, N., Kuffner, R., Heiser, L., Costello, J., Gustavo Stolovitzky, Norman, T., Hoff, B., and Friend, S. (2016). ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge.

[Kurien and Scofield, 2006] Kurien, B. T. and Scofield, R. H. (2006). Western blotting. *Methods*, 38(4):283–293.

[La Manno et al., 2016] La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R., Toledo, E. M., Villaescusa, J. C., Lönnerberg, P., Ryge, J., Barker, R. A., Arenas, E., and Linnarsson, S. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 167(2):566–580.e19.

[La Manno et al., 2018] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494–498.

[Lambert et al., 2018] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors.

[Landry et al., 2003] Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: The role of alternative promoters in mammalian genomes. *Trends in Genetics*, 19(11):640–648.

[Lashkari et al., 1997] Lashkari, D. A., Derisi, J. L., Mccusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062.

[Lee et al., 2014] Lee, H. K., Lee, H. S., and Moody, S. A. (2014). Neural transcription factors: from embryos to neural stem cells. *Molecules and cells*, 37(10):705–712.

[Lee and Young, 2013] Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease.

[Lee and Rio, 2015] Lee, Y. and Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, 84(1):291–323.

[Leivonen et al., 2009] Leivonen, S.-K., Mäkelä, R., Ostling, P., Kohonen, P., Haapa-Paananen, S., Kleivi, K., Enerly, E., Aakula, A., Hellström, K., Sahlberg, N., Kristensen, V. N., Børresen-Dale, A.-L., Saviranta, P., Perälä, M., and Kallioniemi, O. (2009). Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, 28(44):3926–36.

[Lequin, 2005] Lequin, R. M. (2005). Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clinical chemistry*, 51(12):2415–8.

[Levine et al., 2007] Levine, E., Zhang, Z., Kuhlman, T., and Hwa, T. (2007). Quantitative characteristics of gene regulation by small RNA. *PLoS biology*, 5(9):1998–2010.

[Lewis et al., 2005] Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

[Li and Li, 2018] Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):997.

[Li and Zhang, 2015] Li, Y. and Zhang, Z. (2015). Computational Biology in microRNA. *Wiley Interdisciplinary Reviews: RNA*, 6(4):435–452.

[Lim et al., 2005] Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773.

[Lin et al., 2015] Lin, B., He, X., and Ye, J. (2015). A geometric viewpoint of manifold learning. *Applied Informatics*, 2(1).

[Linsley et al., 2007] Linsley, P. S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M. M., Bartz, S. R., Johnson, J. M., Cummins, J. M., Raymond, C. K., Dai, H., Chau, N., Cleary, M., Jackson, A. L., Carleton, M., and Lim, L. (2007). Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Molecular and cellular biology*, 27(6):2240–52.

[Lloyd, 1982] Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Lopez et al., 2018] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

[Love et al., 2015] Love, M. I., Anders, S., Kim, V., and Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res*, 4:1070.

[Lun et al., 2016] Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75.

[Macosko et al., 2015] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.

[Maity et al., 2018] Maity, G., Haque, I., Ghosh, A., Dhar, G., Gupta, V., Sarkar, S., Azeem, I., McGregor, D., Choudhary, A., Campbell, D. R., Kambhampati, S., Banerjee, S. K., and Banerjee, S. (2018). The MAZ transcription factor is a downstream target of the oncoprotein Cyr61/CCN1 and promotes pancreatic cancer cell invasion via CRAF-ERK signaling. *The Journal of biological chemistry*, 293(12):4334–4349.

[Maloy and Hughes, 2013] Maloy, S. and Hughes, K. (2013). *Brenner's Encyclopedia of Genetics*. Elsevier Science.

[Mandel et al., 2008] Mandel, C. R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. *Cellular and Molecular Life Sciences*, 65(7-8):1099–1122.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Aderhold, A., Stolovitzky, G., Bonneau, R., Chen, Y., Cordero, F., Crane, M., Dondelinger, F., Drton, M., Esposito, R., Foygel, R., De La Fuente, A., Gertheiss, J., Geurts, P., Greenfield, A., Grzegorczyk, M., Haury, A. C., Holmes, B., Hothorn, T., Husmeier, D., Huynh-Thu, V. A., Irrthum, A., Karlebach, G., Lèbre, S., De Leo, V., Madar, A., Mani, S., Mordelet, F., Ostrer, H., Ouyang, Z., Pandya, R., Petri, T., Pinna, A., Poultney, C. S., Rezny, S., Ruskin, H. J., Saeys, Y., Shamir, R., Sîrbu, A., Song, M., Soranzo, N., Statnikov, A., Vega, N., Vera-Licona, P., Vert, J. P., Visconti, A., Wang, H., Wehenkel, L., Windhager, L., Zhang, Y.,

and Zimmer, R. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804.

[Martin et al., 2015] Martin, D., Kim, Y. H., Sever, D., Mao, C. A., Haefliger, J. A., and Grapin-Botton, A. (2015). REST represses a subset of the pancreatic endocrine differentiation program. *Developmental biology*, 405(2):316–327.

[Mathelier et al., 2014] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1).

[Matsumoto and Kiryu, 2016] Matsumoto, H. and Kiryu, H. (2016). SCOUP: Probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*, 17(1):1–16.

[Matys, 2006] Matys, V. (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(90001):D108–D110.

[Mauro and Edelman, 2002] Mauro, V. P. and Edelman, G. M. (2002). The ribosome filter hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19):12031–12036.

[Mcadams and Arkin, 1997] Mcadams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–819.

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

[McKenna et al., 2016] McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907.

[Meister, 2013] Meister, G. (2013). Argonaute proteins: Functional insights and emerging roles.

[Memczak et al., 2013] Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., Le Noble, F., and Rajewsky, N. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–338.

[Menon, 2018] Menon, V. (2018). Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics*, 17(4):240–245.

[Mereu et al., 2018] Mereu, E., Iacono, G., Guillaumet-Adkins, A., Moutinho, C., Lunazzi, G., Santos, C., Miguel-Escalada, I., Ferrer, J., Real, F. X., Gut, I., and Heyn, H. (2018). matchSCore: Matching Single-Cell Phenotypes Across Tools and Experiments. *bioRxiv*, page 314831.

[Metzker, 2005] Metzker, M. L. (2005). Emerging technologies in DNA sequencing.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1):31–46.

[Mieczkowski et al., 2016] Mieczkowski, J., Cook, A., Bowman, S. K., Mueller, B., Alver, B. H., Kundu, S., Deaton, A. M., Urban, J. A., Larschan, E., Park, P. J., Kingston, R. E., and Tolstorukov, M. Y. (2016). MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications*, 7(1):11485.

[Miska et al., 2007] Miska, E. A., Alvarez-Saavedra, E., Abbott, A. L., Lau, N. C., Hellman, A. B., McGonagle, S. M., Bartel, D. P., Ambros, V. R., and Horvitz, H. R. (2007). Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS genetics*, 3(12):2395–2403.

[Moon et al., 2018] Moon, K. R., Stanley, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:36–46.

[Mukherji et al., 2011] Mukherji, S., Ebert, M. S., Zheng, G. X. Y., Tsang, J. S., Sharp, P. A., and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. *Nature genetics*, 43(9):854–9.

[Mukhtar and Taylor, 2018] Mukhtar, T. and Taylor, V. (2018). Untangling Cortical Complexity During Development. *Journal of experimental neuroscience*, 12.

[Mullis et al., 1986] Mullis, K. B., Erlich, H. A., Arnheim, N., Horn, G. T., Saiki, R. K., and Scharf, S. J. (1986). Process for amplifying, detecting, and/or-cloning nucleic acid sequences.

[Munsky et al., 2015] Munsky, B., Fox, Z., and Neuert, G. (2015). Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods (San Diego, Calif.)*, 85:12–21.

[Muraro et al., 2016] Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3.

[Nagano et al., 2013] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

[Nam et al., 2014] Nam, J. W., Rissland, O. S., Koppstein, D., Abreu-Goodger, C., Jan, C. H., Agarwal, V., Yildirim, M. A., Rodriguez, A., and Bartel, D. P. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, 53(6):1031–1043.

[Nishimura et al., 2006] Nishimura, W., Kondo, T., Salameh, T., El Khattabi, I., Dodge, R., Bonner-Weir, S., and Sharma, A. (2006). A switch from MafB to MafA expression accompanies differentiation to pancreatic beta-cells. *Developmental biology*, 293(2):526–539.

[Nyayanit and Gadgil, 2015] Nyayanit, D. and Gadgil, C. J. (2015). Mathematical modeling of combinatorial regulation suggests that apparent positive regulation of targets by miRNA could be an artifact resulting from competition for mRNA. *RNA (New York, N.Y.)*, 21(3):307–319.

[Olive et al., 2013] Olive, V., Sabio, E., Bennett, M. J., De Jong, C. S., Biton, A., McGann, J. C., Greaney, S. K., Sodir, N. M., Zhou, A. Y., Balakrishnan, A., Foth, M., Luftig, M. A., Goga, A., Speed, T. P., Xuan, Z., Evan, G. I., Wan, Y., Minella, A. C., and He, L. (2013). A component of the mir-17-92 polycistronic oncomir promotes oncogene-dependent apoptosis. *eLife*, 2013(2).

[Orphanides and Reinberg, 2002] Orphanides, G. and Reinberg, D. (2002). A unified theory of gene expression.

[Osella et al., 2011] Osella, M., Bosia, C., Corá, D., and Caselle, M. (2011). The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS computational biology*, 7(3):e1001101.

[Padovan-Merhar et al., 2015] Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell*, 58(2):339–352.

[Pan et al., 2019] Pan, X., Yang, Y., Xia, C. Q., Mirza, A. H., and Shen, H. B. (2019). Recent methodology progress of deep learning for RNA–protein interaction prediction.

[Paridaen and Huttner, 2014] Paridaen, J. T. and Huttner, W. B. (2014). Neurogenesis during development of the vertebrate central nervous system. *EMBO reports*, 15(4):351–364.

[Patel et al., 2014] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401.

[Paulsson, 2005] Paulsson, J. (2005). Models of stochastic gene expression.

[Pennacchio et al., 2013] Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature reviews. Genetics*, 14(4):288.

[Peterson et al., 2014] Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., and Congdon, C. B. (2014). Common features of microRNA target prediction tools. *Frontiers in Genetics*, 5.

[Phillips, 2008] Phillips, T. (2008). Chromatin remodeling in eukaryotes. *Nature education*.

[Picelli et al., 2013] Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098.

[Poliseno et al., 2010] Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–1038.

[Pollack et al., 1999] Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–46.

[Portales-Casamar et al., 2009] Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. (2009). JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(SUPPL.1).

[Portin, 2002] Portin, P. (2002). Historical Development of the Concept of the Gene. *The Journal of Medicine and Philosophy*, 27(3):257–286.

[Portin and Wilkins, 2017] Portin, P. and Wilkins, A. (2017). The evolving definition of the term "Gene". *Genetics*, 205(4):1353–1364.

[Pray, 2008] Pray, L. (2008). DNA replication and causes of mutation. *Nature Education*.

[Qian et al., 2017] Qian, Y., Li, J., and Xia, S. (2017). ZNF281 Promotes Growth and Invasion of Pancreatic Cancer Cells by Activating Wnt/β-Catenin Signaling. *Digestive diseases and sciences*, 62(8):2011–2020.

[Quapp and Heidrich, 1984] Quapp, W. and Heidrich, D. (1984). Analysis of the concept of minimum energy path on the potential energy surface of chemically reacting systems. *Theoretica Chimica Acta*, 66(3-4):245–260.

[Raj et al., 2008] Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, 5(10):877–879.

[Raj and van Oudenaarden, 2008] Raj, A. and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences.

[Raj et al., 2018] Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., and Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450.

[Rajewsky et al., 2020] Rajewsky, N., Almouzni, G., Gorski, S. A., Aerts, S., Amit, I., Bertero, M. G., Bock, C., Bredenoord, A. L., Cavalli, G., Chiocca, S., Clevers, H., De Strooper, B., Eggert, A., Ellenberg, J., Fernández, X. M., Figlerowicz, M., Gasser, S. M., Hubner, N., Kjems, J., Knoblich, J. A., Krabbe, G., Lichter, P., Linnarsson, S., Marine, J. C., Marioni, J. C., Marti-Renom, M. A., Netea, M. G., Nickel, D., Nollmann, M., Novak, H. R., Parkinson, H., Piccolo, S., Pinheiro, I., Pombo, A., Popp, C., Reik, W., Roman-Roman, S., Rosenstiel, P., Schultze, J. L., Stegle, O., Tanay, A., Testa, G., Thanos, D., Theis, F. J., Torres-Padilla, M. E., Valencia, A., Vallot, C., van Oudenaarden, A., Vidal, M., Voet, T., and Groups., L. C. W. (2020). LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature*, 587(7834):377–386.

[Ramakrishnan, 2002] Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation.

[Regev et al., 2017] Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., and Yosef, N. (2017). The Human Cell Atlas. *Elife*, 6.

[Reinhart et al., 2000] Reinhart, B. J., Slack, F. J., Basson, M., Pasquienelll, A. E., Bettlnger, J. C., Rougvle, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906.

[Ren et al., 2005] Ren, W., Vanden-Eijnden, E., Maragakis, P., and E, W. (2005). Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *The Journal of Chemical Physics*, 123(13):134109.

[Reuter and Mathews, 2010] Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11.

[Riba et al., 2019] Riba, A., Nanni, N. D., Mittal, N., Arhné, E., Schmidt, A., and Zavolan, M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proceedings of the National Academy of Sciences of the United States of America*, 116(30):15023–15032.

[Rio, 2014] Rio, D. C. (2014). Reverse transcription-polymerase chain reaction. *Cold Spring Harbor protocols*, 2014(11):1207–16.

[Risso et al., 2018] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17.

[Rotem et al., 2015] Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172.

[Sanft et al., 2011] Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., and Petzold, L. R. (2011). StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics (Oxford, England)*, 27(17):2457–2458.

[Sanger and Coulson, 1975] Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3).

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

[Schaum et al., 2018] Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., Chen, S., Green, F., Jones, R. C., Maynard, A., Penland, L., Pisco, A. O., Sit, R. V., Stanley, G. M., Webber, J. T., Zanini, F., Baghel, A. S., Bakerman, I., Bansal, I., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Cho, M., Cirolia, G., Conley, S. D., Demers, A., Demir, K., de Morree, A., Divita, T., du Bois, H., Dulgeroff, L. B. T., Ebadi, H., Espinoza, F. H., Fish, M., Gan, Q., George, B. M., Gillich, A., Genetiano, G., Gu, X., Gulati, G. S., Hang, Y., Hosseinzadeh, S., Huang, A., Iram, T., Isobe, T., Ives, F., Kao, K. S., Karnam, G., Kershner, A. M., Kiss, B. M., Kong, W., Kumar, M. E., Lam, J. Y., Lee, D. P., Lee, S. E., Li, G., Li, Q., Liu, L., Lo, A., Lu, W. J., Manjunath, A., May, K. L., May, O. L., McKay, M., Metzger, R. J., Mignardi, M., Min, D., Nabhan, A. N., Ng, K. M., Noh, J., Patkar, R., Peng, W. C., Puccinelli, R., Rulifson, E. J., Sikandar, S. S., Sinha, R., Szade, K., Tan, W., Tato, C., Tellez, K., Travaglini, K. J., Tropini, C., Waldburger, L., van Weele, L. J., Wosczyna, M. N., Xiang, J., Xue, S., Youngyunpipatkul, J., Zardeneta, M. E., Zhang, F., Zhou, L., Castro, P., Croote, D., DeRisi, J. L., Kuo, C. S., Lehallier, B., Nguyen, P. K., Tan, S. Y., Wang, B. M., Yousef, H., Beachy, P. A., Chan, C. K., Huang, K. C., Weinberg, K., Wu, S. M., Barres, B. A., Clarke, M. F., Kim, S. K., Krasnow, M. A., Nusse, R., Rando, T. A., Sonnenburg, J., and Weissman, I. L. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–70.

[Schmiedel et al., 2015] Schmiedel, J. M., Klemm, S. L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D. S., and Van Oudenaarden, A. (2015). Gene expression. MicroRNA control of protein expression noise. *Science (New York, N.Y.)*, 348(6230):128–131.

[Schoenfelder and Fraser, 2019] Schoenfelder, S. and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control.

[Schulze and Downward, 2001] Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*, 3(8):E190–E195.

[Selbach et al., 2008] Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63.

[Setty et al., 2016] Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'Er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):637–645.

[Shahrezaei and Swain, 2008] Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45):17256–17261.

[Shalek et al., 2013] Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240.

[Sheppard et al., 2008] Sheppard, D., Terrell, R., and Henkelman, G. (2008). Optimization methods for finding minimum energy paths. *Journal of Chemical Physics*, 128(13).

[Shin et al., 2015] Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Enikolopov, G., Nauen, D. W., Christian, K. M., Ming, G. L., and Song, H. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell*, 17(3):360–372.

[Shlyueva et al., 2014] Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286.

[Si et al., 2015] Si, J., Cui, J., Cheng, J., and Wu, R. (2015). Computational prediction of RNA-binding proteins and binding sites.

[Smallwood et al., 2014] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, 11(8):817–820.

[Smith and Valcárcel, 2000] Smith, C. W. J. and Valcárcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control.

[Southern, 1975] Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3).

[Spanjaard et al., 2018] Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nature Biotechnology*, 36(5):469–473.

[Stegle et al., 2015] Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145.

[Stuart and Satija, 2019] Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272.

[Suzuki et al., 2009] Suzuki, H., Forrest, A. R. R., van Nimwegen, E., Daub, C. O., Balwierz, P. J., Irvine, K. M., Lassmann, T., Ravasi, T., Hasegawa, Y., de Hoon, M. J. L., Katayama, S., Schroder, K., Carninci, P., Tomaru, Y., Kanamori-Katayama, M., Kubosaki, A., Akalin, A., Ando, Y., Arner, E., Asada, M., Asahara, H., Bailey, T., Bajic, V. B., Bauer, D., Beckhouse, A. G., Bertin, N., Björkegren, J., Brombacher, F., Bulger, E., Chalk, A. M., Chiba, J., Cloonan, N., Dawe, A., Dostie, J., Engström, P. G., Essack, M., Faulkner, G. J., Fink, J. L., Fredman, D., Fujimori, K., Furuno, M., Gojobori, T., Gough, J., Grimmond, S. M., Gustafsson, M., Hashimoto, M., Hashimoto, T., Hatakeyama, M., Heinzel, S., Hide, W., Hofmann, O., Hörnquist, M., Huminiecki, L., Ikeo, K., Imamoto, N., Inoue, S., Inoue, Y., Ishihara, R., Iwayanagi, T., Jacobsen, A., Kaur, M., Kawaji, H., Kerr, M. C., Kimura, R., Kimura, S., Kimura, Y., Kitano, H., Koga, H., Kojima, T., Kondo, S., Konno, T., Krogh, A., Kruger, A., Kumar, A., Lenhard, B., Lennartsson, A., Lindow, M., Lizio, M., Macpherson, C., Maeda, N., Maher, C. A., Maqungo, M., Mar, J., Matigian, N. A., Matsuda, H., Mattick, J. S., Meier, S., Miyamoto, S., Miyamoto-Sato, E., Nakabayashi, K., Nakachi, Y., Nakano, M., Nygaard, S., Okayama, T., Okazaki, Y., Okuda-Yabukami, H., Orlando, V., Otomo, J., Pachkov, M., Petrovsky, N., Plessy, C., Quackenbush, J., Radovanovic, A., Rehli, M., Saito, R., Sandelin, A., Schmeier, S., Schönbach, C., Schwartz, A. S., Semple, C. A., Sera, M., Severin, J., Shirahige, K., Simons, C., St. Laurent, G., Suzuki, M., Suzuki, T., Sweet, M. J., Taft, R. J., Takeda, S., Takenaka, Y., Tan, K., Taylor, M. S., Teasdale, R. D., Tegnér, J., Teichmann, S., Valen, E., Wahlestedt, C., Waki, K., Waterhouse, A., Wells, C. A., Winther, O., Wu, L., Yamaguchi, K., Yanagawa, H., Yasuda, J., Zavolan, M., Hume, D. A., Arakawa, T., Fukuda, S., Imamura, K., Kai, C.,

Kaiho, A., Kawashima, T., Kawazu, C., Kitazume, Y., Kojima, M., Miura, H., Murakami, K., Murata, M., Ninomiya, N., Nishiyori, H., Noma, S., Ogawa, C., Sano, T., Simon, C., Tagami, M., Takahashi, Y., Kawai, J., Hayashizaki, Y., St Laurent, G., Suzuki, M., Suzuki, T., Sweet, M. J., Taft, R. J., Takeda, S., Takenaka, Y., Tan, K., Taylor, M. S., Teasdale, R. D., Tegnér, J., Teichmann, S., Valen, E., Wahlestedt, C., Waki, K., Waterhouse, A., Wells, C. A., Winther, O., Wu, L., Yamaguchi, K., Yanagawa, H., Yasuda, J., Zavolan, M., Hume, D. A., Arakawa, T., Fukuda, S., Imamura, K., Kai, C., Kaiho, A., Kawashima, T., Kawazu, C., Kitazume, Y., Kojima, M., Miura, H., Murakami, K., Murata, M., Ninomiya, N., Nishiyori, H., Noma, S., Ogawa, C., Sano, T., Simon, C., Tagami, M., Takahashi, Y., Kawai, J., and Hayashizaki, Y. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics*, 41(5):553–562.

[Svensson, 2020] Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*.

[Svensson et al., 2017] Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4):381–387.

[Svensson and Pachter, 2018] Svensson, V. and Pachter, L. (2018). RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq.

[Svensson et al., 2018] Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604.

[Swarts et al., 2014] Swarts, D. C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R. F., Koonin, E. V., Patel, D. J., and Van Der Oost, J. (2014). The evolutionary journey of Argonaute proteins.

[Takahashi, 2012] Takahashi, K. (2012). Cellular reprogramming–lowering gravity on Waddington's epigenetic landscape. *Journal of cell science*, 125(Pt 11):2553–60.

[Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76.

[Takahashi and Yamanaka, 2016] Takahashi, K. and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. *Nature Reviews Molecular Cell Biology*, 17(3):183–193.

[Tang et al., 2009] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–82.

[Tauran et al., 2019] Tauran, Y., Poulain, S., Lereau-Bernier, M., Danoy, M., Shinohara, M., Segard, B.-D., Kato, S., Kido, T., Miyajima, A., Sakai, Y., Plessy, C., and Leclerc, E. (2019). Analysis of the transcription factors and their regulatory roles during a step-by-step differentiation of induced pluripotent stem cells into hepatocyte-like cells. *Molecular Omics*.

[Taverna et al., 2014] Taverna, E., Götz, M., and Huttner, W. B. (2014). The cell biology of neurogenesis: toward an understanding of the development and evolution of the neocortex. *Annual review of cell and developmental biology*, 30:465–502.

[Thattai, 2016] Thattai, M. (2016). Universal Poisson Statistics of mRNAs with Complex Decay Pathways. *Biophys. J.*, 110(2):301–305.

[Thompson, 2003] Thompson, W. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research*, 31(13):3580–3585.

[Trapnell et al., 2014] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–6.

[Tronick and Hunter, 2016] Tronick, E. and Hunter, R. G. (2016). Waddington, Dynamic Systems, and Epigenetics. *Frontiers in Behavioral Neuroscience*, 10:107.

[Tsuritani et al., 2007] Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K., and Suzuki, Y. (2007). Distinct class of putative "non-conserved" promoters in humans: Comparative studies of alternative promoters of human and mouse genes. *Genome Research*, 17(7):1005–1014.

[Valencia-Sanchez et al., 2006] Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development*, 20(5):515–24.

[Vallejos et al., 2017] Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities.

[Van Der Maaten and Hinton, 2008] Van Der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

[van der Maaten et al., 2009] van der Maaten, L., Postma, E., and van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, 10:66–71.

[van Dijk et al., 2018] van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27.

[Vasudevan et al., 2007] Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science (New York, N.Y.)*, 318(5858):1931–1934.

[Vaucher and Reiher, 2018] Vaucher, A. C. and Reiher, M. (2018). Minimum Energy Paths and Transition States by Curve Optimization. *Journal of Chemical Theory and Computation*, 14(6):3091–3099.

[Verd et al., 2019] Verd, B., Monk, N. A., and Jaeger, J. (2019). Modularity, criticality, and evolvability of a developmental gene regulatory network. *eLife*, 8.

[Vogel et al., 2010] Vogel, C., de Sousa Abreu, R., Ko, D., Le, S., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M., and Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(1):400.

[Waddington, 1936] Waddington, C. H. (1936). How Animals Develop. *Nature*, 137(3459):251–252.

[Waddington, 1940] Waddington, C. H. (1940). Organisers and genes. *Organisers and genes.*

[Waddington, 1942] Waddington, C. H. (1942). The epigenotype. *Endeavour*, 1:18–20.

[Waddington, 1957] Waddington, C. H. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*

[Wagner et al., 2018] Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987.

[Wahl et al., 2009] Wahl, M. C., Will, C. L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine.

[Wang and Bodovitz, 2010] Wang, D. and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology*, 28(6):281–290.

[Wang et al., 2012] Wang, D., Zhang, Z., O'Loughlin, E., Lee, T., Houel, S., O'Carroll, D., Tarakhovsky, A., Ahn, N. G., and Yi, R. (2012). Quantitative functions of Argonaute proteins in mammalian development. *Genes & development*, 26(7):693–704.

[Wang et al., 2021] Wang, H., Yang, Y., Liu, J., and Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. *Nature Reviews Molecular Cell Biology 2021 22:6*, 22(6):410–424.

[Wang, 2014] Wang, X. (2014). Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics (Oxford, England)*, 30(10):1377–1383.

[Wang et al., 2013] Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X., and Liu, H. (2013). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Developmental cell*, 25(1):69–80.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[Ward, 1963] Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Wee et al., 2012] Wee, L. M., Flores-Jasso, C. F., Salomon, W. E., and Zamore, P. D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, 151(5):1055–67.

[Weirauch et al., 2014] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J. C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F. Y., Ratsch, G., Larrondo,

L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443.

[Welch et al., 2016] Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, 17(1).

[Welch et al., 2019] Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17.

[Wheeler et al., 2018] Wheeler, E. C., Van Nostrand, E. L., and Yeo, G. W. (2018). Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. *Wiley Interdisciplinary Reviews: RNA*, 9(1).

[Wightman et al., 1993] Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862.

[Wolf et al., 2015] Wolf, L., Silander, O. K., and van Nimwegen, E. (2015). Expression noise facilitates the evolution of gene regulation. *eLife*, 4.

[Xia et al., 1998] Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–35.

[Xie et al., 2004] Xie, H., Ye, M., Feng, R., and Graf, T. (2004). Stepwise Reprogramming of B Cells into Macrophages. *Cell*, 117(5):663–676.

[Xuan and Sussel, 2016] Xuan, S. and Sussel, L. (2016). GATA4 and GATA6 regulate pancreatic endoderm identity through inhibition of hedgehog signaling. *Development (Cambridge, England)*, 143(5):780–786.

[Yaghoobi et al., 2012] Yaghoobi, H., Haghipour, S., Hamzeiy, H., and Asadi-Khiavi, M. (2012). A review of modeling techniques for genetic regulatory networks.

[Yan et al., 2016] Yan, D., Wang, H. W., Bowman, R. L., and Joyce, J. A. (2016). STAT3 and STAT6 Signaling Pathways Synergize to Promote Cathepsin Secretion from Macrophages via IRE1α Activation. *Cell Reports*, 16(11):2914–2927.

[Yao et al., 2017] Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A. R., Martinez, R. A., Furchtgott, L., Mulholland, H., Bort, S., Fuqua, M. A., Gregor, B. W., Hodge, R. D., Jayabalu, A., May, R. C., Melton, S., Nelson, A. M., Ngo, N. K., Shapovalova, N. V., Shehata, S. I., Smith, M. W., Tait, L. J., Thompson,

C. L., Thomsen, E. R., Ye, C., Glass, I. A., Kaykas, A., Yao, S., Phillips, J. W., Grimley, J. S., Levi, B. P., Wang, Y., and Ramanathan, S. (2017). A Single-Cell Roadmap of Lineage Bifurcation in Human ESC Models of Embryonic Brain Development. *Cell Stem Cell*, 20(1):120–134.

[Yeung et al., 2018] Yeung, J., Mermet, J., Jouffe, C., Marquis, J., Charpagne, A., Gachon, F., and Naef, F. (2018). Transcription factor activity rhythms and tissue-specific chromatin interactions explain circadian gene expression across organs. *Genome Research*, 28(2):182–191.

[Yu et al., 2006] Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X. S. (2006). Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–1603.

[Yuan et al., 2017] Yuan, G. C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., Quackenbush, J., Saadatpour, A., Schroeder, T., Shivdasani, R., and Tirosh, I. (2017). Challenges and emerging directions in single-cell analysis. *Genome Biology*, 18(1):1–8.

[Yue et al., 2009] Yue, D., Liu, H., and Huang, Y. (2009). Survey of Computational Algorithms for MicroRNA Target Prediction. *Current Genomics*, 10(7):478–492.

[Zappia et al., 2018] Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, 14(6):e1006245.

[Zavolan et al., 2002] Zavolan, M., Van Nimwegen, E., and Gaasterland, T. (2002). Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Research*, 12(9):1377–1385.

[Zeisel et al., 2015] Zeisel, A., Muñ-Manchado, A. B., Codeluppi, S., Lönnerberg, P., Manno, G. L., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., and Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142.

[Zhao et al., 2013] Zhao, H., Yang, Y., and Zhou, Y. (2013). Prediction of RNA binding proteins comes of age from low resolution to high resolution.

[Ziegenhain et al., 2018] Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I., and Enard, W. (2018). Quantitative single-cell transcriptomics. *Briefings in functional genomics*, 17(4):220–232.

[Ziegenhain et al., 2017] Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and

Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643.e4.

Jérémie Breda

1e Virulystraat 3A
3022 ZL Rotterdam
Netherlands
NL +31 61 329 26 23
CH +41 77 426 92 65
jeremiebreda@disroot.org
@jeremiebreda
jmbreda

Education

2014–2020 **Ph.D. in Computational Biology**, *Biozentrum, Universität Basel*, Basel, Switzerland.
· Thesis: *Model-driven analysis of gene expression control.*
· Supervisor: Prof. Dr. Erik van Nimwegen
· Grade: *summa cum laude*

2011–2014 **Master of Science in Physics**, *EPFL*, Lausanne, Switzerland.
· Thesis : *Optimisation of a biophysical microRNA-mRNA interaction model.*
· Supervisor: Prof. Dr. Paolo De Los Rios
· Grade: 6/6

2008–2011 **Bachelor of Science in Physics**, *EPFL*, Lausanne, Switzerland.

Experience

2021-2022 **Postdoctoral researcher**, *Netherlands Cancer Institute (NKI)*, Amsterdam, Netherlands.
Bas van Steensel Lab

2020 **Postdoctoral researcher**, *Biozentrum, Universität Basel*, Basel, Switzerland.
Erik van Nimwegen Lab

2015–2018 **Teaching assistant**, *Universität Basel*, Basel, Switzerland.
· Computational Biology I: Bayesian probability theory
· Introduction to Python programming language

2012–2013 **High school substitute physics teacher**, *Collège de Genève*, Geneva, Switzerland.
Preparing the physics course, teach and organize the work in class.

2009–2012 **Assistant in Math & Physics**, *EPFL*, Lausanne, Switzerland.
· Teaching assistant in the ex cathedra lecture and tutor in the exercise session.

Awards and conferences

2021 **The „Gottfried Schatz“ PhD Student Prize of the Biozentrum, University of Basel**, *Winner*, Basel, Switzerland.

2020 **Bermuda Principles 2020 Impact on Transcriptomics**, *Speaker*, Bermuda.

2019 **Emory Theoretical Biophysics Workshop**, *Participant*, Emory University, Georgia, USA.

2018 **23rd Annual Meeting of the RNA Society**, *Speaker*, UC Berkeley, California, USA.

2017 **Basel Computational Biology Conference**, *Tutorial organiser*, Basel, Switzerland.

2014 **Fellowship for Excellence**, *Awardee*, International PhD Program, Basel, Switzerland.

Other skills and interests

Programming	C++, MatLab, Python, Bash
Languages	French (Mother tongue), English (C2), German (B2)
Music	Drums, Saxophone and Jazz education.
Sport	Climbing, Ski touring and Mountaineering.

Publications

Tanzila Mukhtar, Jeremie Breda, Marcelo Boareto, Pascal Grobecker, Zahra Karimaddini, Alice Grison, Katja Eschbach, Ramakrishnan Chandrasekhar, Swen Vermeul, Michal Okoniewski, Mikhail Pachkov, Suzana Atanasoski, Christian Beisel, Dagmar Iber, Erik van Nimwegen, and Verdon Taylor. Temporal and sequential transcriptional dynamics define lineage shifts in corticogenesis. *bioRxiv*, page 2022.02.10.479992, feb 2022.

Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 39(8):1008–1016, apr 2021.

Tanzila Mukhtar, Jeremie Breda, Alice Grison, Zahra Karimaddini, Pascal Grobecker, Dagmar Iber, Christian Beisel, Erik van Nimwegen, and Verdon Taylor. Tead transcription factors differentially regulate cortical development. *Scientific reports*, 10(1), dec 2020.

Andrzej J Rzepiela, Souvik Ghosh, Jeremie Breda, Arnau Vina-Vilaseca, Afzal P Syed, Andreas J Gruber, Katja Eschbach, Christian Beisel, Erik van Nimwegen, and Mihaela Zavolan. Single-cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction. *Molecular Systems Biology*, 14(8):e8266, aug 2018.

Saeed Omidi, Mihaela Zavolan, Mikhail Pachkov, Jeremie Breda, Severin Berger, and Erik van Nimwegen. Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. *PLOS Computational Biology*, 13(7):e1005176, jul 2017.

Jeremie Breda, Andrzej J Rzepiela, Rafal Gumienny, Erik van Nimwegen, and Mihaela Zavolan. Quantifying the strength of miRNA-target interactions. *Methods (San Diego, Calif.)*, 85, apr 2015.